

**Automated Knowledge Discovery from Functional Magnetic Resonance Images using  
Spatial Coherence**

by

**Pinaki S. Mitra**

B. Tech., Indian Institute of Technology, Kharagpur, 1983

M.S., Virginia Tech, 1987

Submitted to the Graduate Faculty of  
School of Medicine in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy

University of Pittsburgh

2006

UNIVERSITY OF PITTSBURGH

School of Medicine

This dissertation was presented

by

Pinaki S. Mitra

It was defended on

July 26, 2006

and approved by

William F. Eddy Ph.D.

Professor, Department of Statistics, Carnegie Mellon University

George D. Stetten M.D. Ph.D.

Associate Professor, Department of Bioengineering, University of Pittsburgh

Brian E. Chapman Ph.D.

Assistant Professor, Department of Biomedical Informatics, University of Pittsburgh

Gregory F. Cooper M.D. Ph.D.

Associate Professor, Department of Biomedical Informatics, University of Pittsburgh

Dissertation Advisor: Vanathi Gopalakrishnan Ph.D.

Assistant Professor, Department of Biomedical Informatics, University of Pittsburgh

Copyright © by Pinaki S. Mitra

2006

# **Automated Knowledge Discovery from Functional Magnetic Resonance Images using Spatial Coherence**

Pinaki S. Mitra

University of Pittsburgh, 2006

Functional Magnetic Resonance Imaging (fMRI) has the potential to unlock many of the mysteries of the brain. Although this imaging modality is popular for brain-mapping activities, clinical applications of this technique are relatively rare. For clinical applications, classification models are more useful than the current practice of reporting loci of neural activation associated with particular disorders. Also, since the methods used to account for anatomical variations between subjects are generally imprecise, the conventional voxel-by-voxel analysis limits the types of discoveries that are possible. This work presents a classification-based framework for knowledge discovery from fMRI data. Instead of voxel-centric knowledge discovery, this framework is segment-centric, where *functional segments* are clumps of voxels that represent a functional unit in the brain. With simulated activation images, it is shown that this segment-based approach can be more successful for knowledge discovery than conventional voxel-based approaches. The spatial coherence principle refers to the homogeneity of behavior of spatially contiguous voxels. Auto-threshold Contrast Enhancing Iterative Clustering (ACEIC) – a new algorithm based on the spatial coherence principle is presented here for functional segmentation. With benchmark data, it is shown that the ACEIC method can achieve higher segmentation accuracy than Probabilistic Independent Component Analysis – a popular method used for fMRI data analysis. The spatial coherence principle can also be exploited for voxel-centric image-classification problems. Spatially Coherent Voxels (SCV) is a new feature selection method that uses the spatial coherence principle to eliminate features that are unlikely to be useful for classification. For a Substance Use Disorder dataset, it is demonstrated that feature selection with SCV can achieve higher classification accuracies than conventional feature selection methods.

## TABLE OF CONTENTS

<b>ACKNOWLEDGMENTS .....</b>	<b>XVI</b>
<b>GLOSSARY.....</b>	<b>XVII</b>
<b>1.0 INTRODUCTION.....</b>	<b>1</b>
<b>1.1 THE PROBLEM.....</b>	<b>2</b>
<b>1.2 THE APPROACH .....</b>	<b>4</b>
<b>1.2.1 Thesis.....</b>	<b>5</b>
<b>1.3 SIGNIFICANCE.....</b>	<b>6</b>
<b>1.4 DISSERTATION OVERVIEW .....</b>	<b>7</b>
<b>2.0 BACKGROUND .....</b>	<b>8</b>
<b>2.1 IMAGING BRAIN FUNCTION .....</b>	<b>8</b>
<b>2.1.1 Magnetic Resonance Phenomenon .....</b>	<b>9</b>
<b>2.1.2 MRI Pulse Sequences.....</b>	<b>12</b>
<b>2.1.3 Spatial Encoding .....</b>	<b>13</b>
<b>2.1.4 BOLD Contrast .....</b>	<b>16</b>
<b>2.1.5 Limitations of fMRI data .....</b>	<b>19</b>
<b>2.1.5.1 Technological Limitations.....</b>	<b>19</b>
<b>2.1.5.2 Physiological Limitations .....</b>	<b>19</b>
<b>2.2 CONVENTIONAL DATA ANALYSIS PIPELINE.....</b>	<b>20</b>
<b>2.2.1 Experimental Design.....</b>	<b>21</b>
<b>2.2.1.1 Experimental Paradigm .....</b>	<b>21</b>
<b>2.2.1.2 Stimulus Presentation.....</b>	<b>21</b>
<b>2.2.2 Analysis pipeline.....</b>	<b>23</b>
<b>2.2.2.1 Pre-processing steps.....</b>	<b>24</b>
<b>2.2.2.2 Voxel-wise Statistical Analysis.....</b>	<b>26</b>

2.2.2.3	Spatial Normalization.....	29
2.2.2.4	Group map from multiple subjects .....	30
2.2.2.5	Comparison between groups of subjects .....	31
2.3	MACHINE LEARNING METHODS.....	31
2.3.1	Machine Learning Concepts .....	32
2.3.1.1	Features and Attributes .....	32
2.3.1.2	Model selection and validation .....	33
2.3.1.3	Feature construction and selection.....	35
2.3.2	Classical Statistical analysis vs. Machine Learning.....	35
2.3.3	Gaussian Naïve Bayes Classifier.....	36
2.3.4	Artificial Neural Networks.....	38
2.3.5	Support Vector Machine .....	41
2.4	MACHINE LEARNING FROM ACTIVATION MAPS .....	45
2.4.1	Feature construction and selection.....	46
2.4.1.1	PCA .....	46
2.4.1.2	<i>k</i> -best voxels (KBV) .....	48
2.4.2	Motivations for new approaches.....	48
2.5	CONVENTIONAL FUNCTIONAL SEGMENTATION .....	50
2.5.1	Clustering.....	50
2.5.1.1	Distance Metrics for Clustering .....	50
2.5.1.2	Clustering Methods .....	51
2.5.2	Independent Component Analysis (ICA).....	54
2.5.3	Image processing methods .....	56
3.0	ANNOTATED EXAMPLE .....	58
3.1	CLINICAL APPLICATION: SUBSTANCE USE DISORDER.....	58
3.1.1	Experimental Paradigm .....	59
3.1.2	Data collection protocol.....	61
3.2	DATA ANALYSIS.....	62
3.2.1	Pre-processing .....	62
3.2.2	Statistical Analysis of Single Brain.....	62
3.2.3	Group Analysis.....	63

<b>4.0</b>	<b>KDSF FRAMEWORK .....</b>	<b>64</b>
<b>4.1</b>	<b>DESCRIPTION OF KDSF .....</b>	<b>66</b>
4.1.1	Functional Segmentation.....	66
4.1.2	ROA Registration.....	68
4.1.3	Feature Construction.....	69
4.1.4	Automated Knowledge Discovery .....	70
<b>4.2</b>	<b>EVALUATION OF KDSF .....</b>	<b>71</b>
4.2.1	Baseline Data .....	72
4.2.2	Model of Region of Activation (ROA).....	74
4.2.3	Generative Model.....	76
4.2.4	Research Design .....	80
4.2.5	Feature Construction/Selection .....	82
4.2.6	Hypothesis Testing.....	83
4.2.7	KDSf and segmentation accuracy.....	85
4.2.8	Machine Learning Details .....	86
<b>4.3</b>	<b>KDSF RESULTS .....</b>	<b>86</b>
4.3.1	Test-Suite-Size (Without Smoothing).....	86
4.3.1.1	KDSf vs. KBV .....	86
4.3.1.2	KDSf vs. PCA .....	88
4.3.2	Test-Suite-Size (With Smoothing) .....	91
4.3.3	Test-Suite-Size: Effect of Location-variability .....	93
4.3.4	Test-Suite-Size: Effects of Smoothing .....	95
4.3.5	Test-Suite-Size: KDSf with inaccurate segmentation.....	98
4.3.6	Test-Suite-CNR .....	99
4.3.6.1	KDSf vs. KBV .....	99
4.3.6.2	KDSf vs. PCA .....	101
4.3.7	Conclusions from KDSf simulation study .....	102
<b>5.0</b>	<b>SEGMENTATION WITH ACEIC .....</b>	<b>104</b>
<b>5.1</b>	<b>RELATED WORK.....</b>	<b>106</b>
<b>5.2</b>	<b>DETAILS OF ACEIC .....</b>	<b>107</b>
5.2.1	CEIC.....	110

5.2.2	ACEIC.....	113
5.3	EVALUATION OF ACEIC.....	114
5.3.1	Evaluation with Benchmark Data.....	114
5.3.1.1	Datasets HDS1/HDS2 .....	114
5.3.1.2	Results for HDS1/HDS2 .....	115
5.3.2	Application to experimental data .....	118
5.4	DISCUSSION.....	120
6.0	MACHINE LEARNING WITH SCV .....	122
6.1	RELATED WORK.....	123
6.2	METHODS.....	125
6.2.1	Dataset.....	125
6.2.2	Normalization .....	126
6.2.3	Feature Selection.....	126
6.2.4	Spatially Coherent Voxels (SCV) .....	127
6.2.5	Classification .....	129
6.3	RESULTS .....	130
6.4	DISCUSSION.....	133
7.0	CONCLUSIONS AND FUTURE WORK .....	134
7.1	SPECIFIC FINDINGS .....	134
7.2	FUTURE WORK.....	135
	APPENDIX.....	137
	BIBLIOGRAPHY .....	141



## LIST OF TABLES

Table 1. Gyro-magnetic ratio for some nuclei [15]. .....	10
Table 2. Characteristic T1 and T2 parameters for different tissue types [15]. .....	11
Table 3. Parameters for simulated regions of activation (ROA). .....	75
Table 4. Parameters for generative model for a population of activation images. ....	79
Table 5. Relationship between generative model parameters and ROA parameters. The Gaussian distributions for ROA parameters (rows) are controlled by the parameters of the generative model. Note ‘*’ signifies multiplication. ....	79
Table 6. Generative model parameters explored in Test-Suite-Size and Test-Suite-CNR. ....	81
Table 7. Test-Suite-Size (no smoothing): Improvements in classification accuracy of KDSf compared to that of KBV. ....	88
Table 8. Test-Suite-Size (no smoothing): Improvement of classification accuracy of KDSf over PCA. ....	89
Table 9. Test-Suite-Size: Typical distribution of ROA sizes for $sd(\sigma^N)=0$ . ....	91
Table 10. Test-Suite-Size (with spatial smoothing): Accuracies for KDSf, KBV and PCA ( $sd(\sigma^N)=0$ ). ....	93
Table 11. Test-Suite-CNR: Classification accuracy of KDSf compared to that of KBV and PCA. ....	102
Table 12. ND dataset: Leave-one-out classification accuracies for the three feature selection methods – best accuracies achieved over respective parameter spaces (number of components, $k$ , and $\theta$ ). ....	132

## LIST OF FIGURES

Figure 1. Nuclear Magnetic Resonance. a) Precession of the net magnetization vector around the longitudinal magnetic field while recovering from a 90 degree flip to the transverse plane (caused by introduction of an RF pulse). b) The receiver coil in the transverse plane detects the signal from the precessing magnetization vector. c) Decaying echo signal recorded (read out) from the receiver coil. ....	12
Figure 2. Gradient Echo technique. a) All protons are in phase. b) When gradient is applied, some protons spin faster than before (red) and some slow down (blue), leading to loss of signal. c) Reversing the gradient speeds up the previously slowed down protons and vice versa. d) Protons are in phase again and produce a new signal echo. ....	13
Figure 3. Echo Planar Imaging (EPI) a) Pulse sequence for Echo Planar Imaging. b) k-space trajectory as the signal from the receiver coil is sampled over time. ....	16
Figure 4. Decay of transverse magnetization is slower during activation (larger $T2^*$ ) – thus, MR signal amplitude is higher for activated state. ....	17
Figure 5. Hemodynamic response to brief stimulus. a) Resting state. b) Neural activation is associated with higher concentration of oxy-hemoglobin (red circles) in the blood compared to the resting state. c) Idealized hemodynamic response function. ....	18
Figure 6. Comparison of Block (top row) and Event-related (bottom row) task designs. ....	22
Figure 7. Conventional data-processing pipeline for fMRI data .....	24
Figure 8. Example group activation map (thresholded by t-score and overlaid on anatomical image of template brain). ....	31
Figure 9. A node in an Artificial Neural network. ....	39
Figure 10. Artificial Neural Network with one hidden layer with two nodes. ....	40
Figure 11. The margin (dotted line) for SVM separating boundary. ....	42

Figure 12. Penalty terms ( $\hat{U}$ ) for misclassification by SVM. ....	43
Figure 13. Schematic of spatial ICA of fMRI data.....	55
Figure 14. Main cortical areas ('functional units') involved in saccadic control (adapted from [69]).....	60
Figure 15. The VGS task alternates between 30 seconds blocks of fixation (a) and pro-saccade (b and c). During the pro-saccade task, the subject's gaze moves towards the dot when the probe is presented. During the pro-saccade block, the probe randomly alternates between Probe 1 and Probe 2. ....	60
Figure 16. Difference between pro-saccade and anti-saccade conditions. a) During the pro-saccades condition (cued by green cross), the subject is instructed to look towards the dot when it appears. b) During the anti-saccades condition, the subject is instructed to look away from the location of the dot (inhibit reflexive pro-saccadic eye movement). In both cases, the probe randomly alternates between Probe 1 and Probe 2 (see Figure 15). ....	61
Figure 17. Example activation image for the VGS task (axial slices of 3D image volume). For each voxel, the z-score is represented as a heat-map.....	62
Figure 18. Hypothetical differences in activation patterns between two populations of subjects.	63
Figure 19. Outline of the KDSf framework for Knowledge Discovery. ....	65
Figure 20. Functional segmentation based upon similarity of adjacent time-courses. The top row represents the time-series of functional images. Timecourses for pixels within the Region of Activation (e.g. pixels $i$ and $j$ in bottom image) are similar compared to the timecourses outside the ROA (e.g. pixel $m$ ). ....	67
Figure 21. ROA registration determines the correspondence of ROAs across subjects. The ROAs from a set of segmented images (top row) are labeled (colored) by the registration process (bottom row) – each color corresponds to a functional unit. ....	69
Figure 22. KDSf feature construction: The machine learning table is constructed from attributes of functional units (e.g. size and average t-score). The functional units are labeled with colors.	70
Figure 23. Schematic for creation of synthetic activation images. Simulated activation (on-off pattern) is added to the time-series for the set of pixels inside the region of activation (top arrow). ....	74

Figure 24. A randomly generated 2D Gaussian function is thresholded to generate random shapes for regions of activation. The pixels in the 2D Gaussian (top, left) for which the value exceeds half of the maximum value (vertical arrow) are retained in the region of activation (top, right).	75
Figure 25. Example simulated activation images from datasets created with generative models. a) Between-group differences in ROA-size. b) Between-group differences in activation-level (indicated by brightness of the ROA). In both cases, ROAs exhibit between-subject location-variability and size-variability.	76
Figure 26. Steps for generation of synthetic dataset. Multiple activation images are created from a generative model.	77
Figure 27. Overview of accuracy computations (for KDSf, PCA and KBV) from a synthetic dataset created with a generative model.	84
Figure 28. Test-Suite-Size (no smoothing): Results for Hypothesis 4.1. The validity of the hypothesis $\text{Mean}(\text{Acc}_{\text{KDSf}}) > \text{Mean}(\text{Acc}_{\text{KBV}})$ , for different conditions is shown in the top row. The bottom row shows the improvement in accuracy of KDSf over KBV.	87
Figure 29. Test-Suite-Size (no smoothing): Results for Hypothesis 4.2. The validity of the hypothesis $\text{Mean}(\text{Acc}_{\text{KDSf}}) > \text{Mean}(\text{Acc}_{\text{PCA}})$ for different conditions is shown in the top row. The bottom row shows the improvement in accuracy of KDSf over PCA.	89
Figure 30. Differentially activated pixels. a) Examples of regions of activation for ‘disease’ images. b) Examples of regions of activation for ‘normal’ images c) Overlay of regions of activation shows the subset of pixels that are consistently activated in ‘disease’ but not in ‘normal’.	90
Figure 31. Test-Suite-Size: Effect of spatial smoothing on Hypothesis 4.1. The validity of the hypothesis $\text{Mean}(\text{Acc}_{\text{KDSf}}) > \text{Mean}(\text{Acc}_{\text{KBV}})$ , for different conditions is shown in the top row. The bottom row shows the improvement in accuracy of KDSf over KBV.	92
Figure 32. Test-Suite-Size: Effect of spatial smoothing on Hypothesis 4.2. The validity of the hypothesis $\text{Mean}(\text{Acc}_{\text{KDSf}}) > \text{Mean}(\text{Acc}_{\text{PCA}})$ , for different conditions is shown in the top row. The bottom row shows the improvement in accuracy of KDSf over PCA.	92
Figure 33. Test-Suite-Size: Feature selection in the absence of location-variability ( $\sigma^{\text{D/N}}=1.2$ , $\text{sd}(\sigma^{\text{N}})=0$ , $\text{sd}(\mu)=0$ , and no smoothing). a) Heat-map image showing spatial distribution of ROA for the ‘normal’ group. b) Heat-map image for ‘disease’ group. c) Eigenimage 2 (as heat-map)	

identified by PCA. d) Interestingness values as heat-map: locations of the ‘best’ voxel features for KBV are shown in red.....	94
Figure 34. Test-Suite-Size: Feature selection with location-variability ( $\sigma^{D/N}=1.2$ , $sd(\sigma^N)=0$ , $sd(\mu)=2$ and no smoothing). a) Heat-map image showing spatial distribution of ROA for the ‘normal’ group. b) Heat-map image for ‘disease’ group. c) Eigenimage 4 (as heat-map) identified by PCA. d) Interestingness values as heat-map: locations of the ‘best’ voxel features for KBV are shown in red.....	95
Figure 35. Smoothing an activation image (SPM). Heat-map representation of a simulated activation image before (left) and after smoothing (right) with a Gaussian kernel (FWHM=8). Higher values are shown with ‘warmer’ colors. ....	96
Figure 36. Test-Suite-Size: Effect of smoothing on feature construction in the presence of location-variability ( $\sigma^{D/N}=1.2$ , $sd(\sigma^N)=0$ , $sd(\mu)=2$ and FWHM=8). a) Eigenimage 1 from PCA. b) Eigenimage 2 from PCA. c) Eigenimage 4 from PCA. d) Locations of the ‘best’ voxel features for KBV (images smoothed prior to ‘interestingness’ calculations). ....	96
Figure 37. Test-Suite-Size: Effect of smoothing (FWHM=8) on activation image (SPM) for one ‘normal’ subject. The image is shown in profile for $y=32$ . The peak is the region of activation. ....	97
Figure 38. Effect of smoothing (FWHM=8) on activation image (SPM) for one ‘disease’ subject. The image is shown in profile for $y=32$ . The peak is the region of activation (note that ROA is wider for the ‘disease’ subject, compared to Figure 37).....	98
Figure 39. Test-Suite-Size: Results for Hypothesis 4.3. The validity of the hypothesis, $Mean(Acc_{KDSf}) > Mean(Acc_{KDiSf})$ , for different conditions is shown in the top row. The bottom row shows the improvement in accuracy of KDSf over KDiSf.....	99
Figure 40. Test-Suite-CNR: Results for Hypothesis 4.1 (‘size’ variability, $sd(\sigma^N)=0.1$ ). The validity of the hypothesis $Mean(Acc_{KDSf}) > Mean(Acc_{KBV})$ , for different conditions is shown in the top row. The bottom row shows the improvement in accuracy of KDSf over KBV.....	100
Figure 41. Test-Suite-CNR: Results for Hypothesis 4.2 (‘size’ variability, $sd(\sigma^N)=0.1$ ). The validity of the hypothesis $Mean(Acc_{KDSf}) > Mean(Acc_{PCA})$ for different conditions is shown in the top row. The bottom row shows the improvement in accuracy of KDSf over PCA. ....	101
Figure 42. Test-Suite-Size: Effect of smoothing on voxel-based feature construction in the presence of location-variability ( $\sigma^{D/N}=2$ , $sd(\sigma^N)=0.1$ , $sd(\mu)=1$ ). a) Without smoothing (FWHM=0), locations of ‘interesting’ voxels correctly suggest a ‘size’ difference between	

groups. b) With smoothing (FWHM=8), locations of ‘interesting’ voxels incorrectly suggest that the groups differ in strengths of activation. ....	103
Figure 43. Under-segmentation problem associated with contrast maximization with greedy agglomeration. The global maximum may be away from the correct solution (at 25 pixels). ....	109
Figure 44. Algorithm CEIC – Contrast Enhancing Iterative Clustering.....	111
Figure 45. Convergence of CEIC (iterations 8 and 18 are the same). The contrasts for the candidate region-definitions are also shown.....	112
Figure 46. Automated threshold selection by the ACEIC method. The optimal thresholds are shown with filled symbols. ....	113
Figure 47. a) Locations of artificial regions of activation in HDS1 (upper square) and HDS2 (lower square). b) Locations of pre-existing structures in the baseline image are highlighted in the heat-map image, where ‘warmer’ colors indicate the presence of strong periodic components in the timecourses. ....	115
Figure 48. Comparison of accuracies of ACEIC and MELODIC for dataset HDS1. ....	116
Figure 49. Examples of MELODIC solutions. a) Thresholded component map (showing z-scores) for CNR=0.66 (note false-positives). b) Thresholded Component map for CNR=3 (correct segmentation solution, no false positives). c) Timecourse associated with component shown for CNR=0.66. d) Timecourse associated with component shown for CNR=3.00.....	117
Figure 50. Comparison of accuracies of ACEIC and MELODIC for dataset HDS2. ....	118
Figure 51. ACEIC segmentation of fMRI data from saccade experiment. a) T-test activation map showing regions of activation. b) A subset of segments identified by ACEIC. c) Timecourses associated with ACEIC segments. ....	119
Figure 52. Example of spatially coherent voxels. a) Heat-map representation of activation values for a coherent ROI. b) For the coherent ROI, average accuracy (ACA) stays above threshold as the ROI is grown one voxel at a time. c) Activation values for less coherent ROI. d) Average accuracy quickly falls below threshold (0.9). ....	128
Figure 53. ND dataset: Leave-one-out classification accuracies for SVM models with feature selection by KBV and feature refinement by SCV (with the same $k$ ). ....	131
Figure 54. ND dataset: Leave-one-out classification accuracies for PCA-based feature selection, with different number of components in the model.....	132

Figure 55. Location of spatially coherent voxels that were used in the best SCV-based classification configuration ( $k=100$ voxels, $a_R=0.9$ , and $\theta=20 \text{ mm}^3$ ).....	132
Figure 56. Greedy ROA registration algorithm. ....	140

## **ACKNOWLEDGMENTS**

I would like to thank my advisor Vanathi Gopalakrishnan Ph.D. for her guidance and help with my research. I thank William Eddy Ph.D. of Carnegie Mellon University for the opportunity to be associated with the Imaging Group at SNOW (Statistics NOrth West). I thank my committee members, Gregory Cooper M.D. Ph.D., George Stetten M.D. Ph.D., and Brian Chapman Ph.D. for their steady support and guidance.

Special thanks to Rebecca McNamee Ph.D. for helpful discussions and for the Substance Use Disorder dataset used here.

I would like to thank Cleat Szczepaniak, Rose Ann Thomas, Toni Porterfield, Joseph Cummings and William Milberry for making the DBMI experience so enjoyable.

Finally, I would like to acknowledge the financial support of the National Library of Medicine (Medical Informatics Training Grant number 5 T15 LM/DE07059).



## GLOSSARY

Accuracy: For a classification model, accuracy measures the degree to which the classes predicted for subjects (e.g. normal vs. disease) match the true classes of subjects. Accuracy is a number between 0 and 1 that is computed from counts of correct and incorrect classification.

		True Class	
		Disease	Normal
Prediction	Disease	True Positive (TP)	False Positive (FP)
	Normal	False Negative (FN)	True Negative (TN)

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}}$$

ACEIC: Auto-threshold Contrast Enhancing Iterative Clustering, an algorithm for segmentation of fMRI images.

Activation: Neural activity associated with performance of a task.

Activation Level: Strength of activation as reflected by the increase in signal strength induced by neural activation.

Activation Image: See SPM.

BOLD Contrast: Blood Oxygenation Level Dependent Contrast. The task-related change in the MR signal is dependent upon the level of blood oxygenation around neural tissue.

Box-car: In a ‘box-car’ design for an fMRI experiment, the subject repeatedly performs some task over a block of time followed by a block of time for rest – this sequence is repeated several times.

Classification Models: Computer programs that can predict the ‘class’ (or group-membership) of subjects (e.g. normal or disease) based upon ‘features’ (or attributes) of subjects.

CNR (Contrast to Noise Ratio): CNR is the ratio of the amplitude of the signal change due to neural activation (BOLD response) and an estimate of the inherent noise-level of the signal. The

noise-level is typically estimated from the standard deviation (over time) of the signal from un-activated voxels within the brain.

Feature Construction: Feature construction is the creation of ‘attributes’ to be used for machine learning. Since it is difficult to use images directly for machine learning, attributes of the images are isolated to represent the information-content of the images. Features can be constructed from individual voxels or from a group of voxels (PCA or segmentation).

Feature Selection: Since the inclusion of irrelevant features in a classification model can reduce the accuracy of the model, a subset of all possible features is selected for inclusion in the classification model.

Functional MRI (fMRI): Functional Magnetic Resonance Imaging scans are a series of images that track changes in blood oxygenation due to neural *activation* associated with performance of a task during the experiment.

Functional Segmentation: Demarcation of Functional Units in a brain image based upon the function of the corresponding brain tissue. This is based upon the working hypothesis that the brain tissues involved in similar function are co-modulated during the experimental task.

Functional Units: Pockets of cortical tissue that specialize in a particular function (e.g. Frontal Eye Fields).

FWHM: Full-Width-at-Half-Maximum specifies the ‘spatial extent’ of a Gaussian function (kernel) used for smoothing.

GNB: Gaussian Naïve Bayes – a machine learning method that employs Bayes rule for classification.

Hemodynamic Response Function (HRF): Local changes in blood flow and oxygenation level over time as a response to brief neural activity. The MR signal reflects these changes.

KBV: *k*-best voxels – a feature selection method that ranks voxels by some criterion and selects the top *k* voxels as features for classification.

KDSf: Knowledge Discovery from Segmented fMRI – a high-level framework for *segment-centric* knowledge discovery from fMRI data.

Knowledge Discovery and Data-mining (KDD): Automated discovery of novel and useful information from large datasets. Here, from a set of fMRI images, the goal is to determine which areas (regions) of the brain are differentially activated between different populations of subjects (e.g. normal vs. disease).

Location-variability: The between-subject variability of exact locations of regions of activation (ROA).

Machine Learning: A set of techniques that predict some attribute (e.g. class) of subjects based upon other attributes of subjects.

MRI: Magnetic Resonance Imaging.

Naïve Bayes: A machine learning method that uses Bayes Rule to classify subjects.

Pixel: Picture cell. A pixel is the elemental unit in a 2D image of the brain. See voxels.

PCA: Principal Component Analysis creates a smaller number of uncorrelated variables from linear combinations of correlated variables.

p-value: Assuming that the null-hypothesis is true (no effect), the p-value is the probability of observing a value for the test-statistic, equally or more surprising (in the direction of the alternative hypothesis) than the value of the test-statistic actually observed.

Region of Interest (ROI): A pocket of pixels/voxels in the image.

Region of Activation (ROA): A pocket of pixels/voxels representing neurons ‘activated’ by performance of a *task*.

Segmentation: Segmentation is the grouping of image elements into clumps (pixels or voxels) based upon some characteristic of the image elements. These clumps of pixels or voxels are called segments.

Segment-centric KD: Knowledge discovery based upon the assumption that functional segments (or functional units in the brain) are comparable across subjects.

Signal-to-noise ratio (SNR): A ratio of the energy-content of the signal of interest (an object in an image) to the energy-content of the signal not of interest (background of image).

Spatial Coherence: The similarity of behavior of spatially contiguous voxels. The homogeneity of evidence within a larger spatial cluster of voxels is more informative than that from individual voxels.

Spatially Coherent Voxels (SCV): Spatially contiguous voxels that present homogeneous evidence for differential activation between groups.

Spatial Normalization: Transformation of images of brains (of different shapes and sizes) to a common co-ordinate system (based on an anatomical template) such that voxel locations in the transformed brain images are approximately comparable.

Statistical Parameter Map (SPM): A model-driven method for analysis of fMRI data where voxel-wise regression is used to estimate the degree to which the timecourse for a voxel conforms to the activation pattern expected by the model. A 3D map of these regression coefficients represents the activation levels at different regions of the brain – thus this 3D image is also referred to as the activation image.

Stimulus: The information presented to the subject during the experimental *task*.

Support Vector Machine (SVM): A classification model that is based upon an optimal separating boundary in feature-space.

Task: During the course of the fMRI experiment, the subject performs one or more tasks while the scanner collects images. A typical task might involve viewing of images and responding with finger taps. The task-history (or task ‘timecourse’) is a record of the tasks performed by the subject over time (including resting times). A task typically consists of multiple *conditions* for comparison purposes.

Testing set: A dataset that is used to test the accuracy of a classification model.

Test-Suite-Size: Synthetic image dataset used for testing KDSf when sizes of ROAs differ between groups.

Test-Suite-CNR: Synthetic image dataset used for testing KDSf when activation levels of ROAs differ between groups.

Timecourse: A time-series of measurements – typically the MR signal strength over the course of an fMRI experiment.

Training set: A dataset that is used to train a classification model. Training refers to learning the optimal parameters of the model from the data.

Voxel: Volume cell. A voxel is the elemental unit in a 3D image of the brain. A voxel represents a physical volume inside the subject’s brain. All the methods described here are applicable to both 2D and 3D images, hence the terms voxel and pixel are used interchangeably.

Voxel-centric KD: Knowledge discovery based on the assumption that voxels are comparable across subjects (after spatial normalization).

## 1.0 INTRODUCTION

The human brain has been described as the most complex system in the known universe. While many mysteries about the brain remain unsolved, the physical architecture of the brain is fairly well mapped. In vivo imaging of the brain is now part of standard clinical practice. While structural abnormalities are routinely imaged for clinical assessment, imaging of functional abnormalities is also slowly becoming a reality. The possibility of in vivo imaging of brain function raises intriguing possibilities for improved understanding and management of disorders such as schizophrenia, substance use disorders, and Alzheimer's disease.

The availability of large numbers of brain images – both structural and functional – provides an unprecedented opportunity for automated knowledge discovery and data-mining from these images. Knowledge discovery may involve discovery of new associations between cognitive dysfunction and characteristics of images, or even creation of automated diagnosis tools that can be used to classify subjects based upon patterns of brain structure [1] or neural activation [2].

Functional Magnetic Resonance Imaging (fMRI) can be used to study neural activation by tracking blood oxygenation level dependent [3] changes to the magnetic resonance signal over time during performance of a task. In an fMRI experiment, for each location (voxel) in the brain, a timecourse (or time-series) captures the signal changes associated with neural activation around that location, delayed and blurred by the hemodynamic response to the neural activation. During the experiment, the subject is requested to perform some task while MR image volumes are collected sequentially. Statistical analysis of this sequence of images provides information about the foci of neural activity underlying the brain processes corresponding to the specific task. Such information about the loci and level of neural activation can be useful for clinical or research purposes. For example, language lateralization can be assessed with fMRI prior to

neurosurgery; or, language-related activation can be compared for two groups of subjects drawn from different populations.

## **1.1 THE PROBLEM**

Several types of data analysis methods are employed for fMRI studies. In brain mapping studies, the goal is the isolation of a consensus region of neural activation associated with a particular function of the brain – statistical analysis methods based upon classical hypothesis rejection are typically used for this purpose. Similar approaches are also used for comparison of activation patterns between two groups of subjects. However, construction of automated clinical tools that can be used for diagnosis and management of cognitive disorders requires the use of statistical and machine learning methods.

Automated knowledge discovery involves the application of machine learning methods to an fMRI dataset to discover clinically useful patterns in the data. The knowledge content in such discovered patterns may be explicit (e.g., dorsolateral prefrontal cortex activation is lower for subjects with Alzheimer’s disease) or the knowledge may be implicit in a classification model (e.g., a program that can discriminate between fMRI images from normal subjects and subjects with Alzheimer’s disease). Automated knowledge discovery from fMRI images entails construction of ‘features’ or attributes based upon the images and searching for classification models constructed from these features that can reliably predict clinically useful outcomes. The discovered models can then be used directly in clinical practice, or they can be inspected manually to extract the discovered ‘knowledge content’.

In this approach to automated knowledge discovery, detection of useful differences between groups of subjects is dependent upon the accuracy of classification models. When patterns of neural activation are indeed different between two groups, the classification models should be able to accurately discriminate between the groups. Feature-construction and feature-selection are critical steps that determine the accuracy of classification models. Current approaches to feature-construction are mainly voxel-based – the voxel is treated as the unit of activation and analysis. However this approach suffers from the following drawbacks:

## **1. Small Datasets**

Since scanner resources are relatively expensive, typical fMRI studies are limited to a handful of subjects. The large image sizes (several thousand voxels in each brain volume) and small numbers of subjects in the studies pose the risk of ‘over-fitting’ with voxel-based classification models – the models may not generalize to subjects not in the original study.

## **2. Inter-subject variability of activation**

A wide variety of morphological differences are observed between brains of different subjects. Also, substantial inter-subject variability is observed for loci of neural activation associated with performance of the same task. The current approach to this problem involves spatial normalization of the individual brain images followed by smoothing of the images prior to voxel-based statistical analysis. In the spatial normalization step, the brain image is morphed into the shape of a common anatomical template so that brains of different shapes and sizes can be compared. Additionally, the images are smoothed to bring anatomical structures into general correspondence. Finally, statistical (or machine learning) techniques are applied to these normalized and smoothed set of voxels for identification of differentially activated regions in the brain. This voxel-based approach to data analysis is not well-suited to handle the inter-subject variability problem – empirical studies have shown that current spatial normalization methods may not lead to sufficient overlap between anatomical structures in images from different subjects [4, 5]. For example, Nieto-Castanon et al. [5] used anatomical markers to manually identify 10 Regions Of Interest (ROI) from perisylvian cortical areas in the temporal and parietal lobes and compared the voxel-overlap between these ROIs across 9 subjects. In this study, nine of the ROIs showed no voxel-overlap across all normalized subjects. Only the largest of the 10 ROIs showed non-zero (less than 5%) overlap across all nine subjects. While spatial smoothing may bring some of the voxels into registration, it also causes loss of spatial resolution – there is no consensus regarding the optimal degree of smoothing. Smoothing can also impact the ability to detect between-group differences in activation levels (see Chapter 4).

## **3. Hypotheses about size of activation**

A more serious problem with voxel-based analysis is that only certain kinds of hypotheses can be tested directly with voxel-based methods. Typically, the hypotheses tested with voxel-based methods are about strengths of activation at a particular voxel location in the brain – hypotheses about different sizes of activated regions cannot be tested directly. While

larger activated regions in a particular group of subjects may be reflected in higher t-scores around the periphery of the consensus region of activation (see Chapter 4), this effect can be unreliable in the presence of inadequate spatial normalization of inter-subject variation in the location of activation.

## 1.2 THE APPROACH

In this work, an alternative to voxel-centric knowledge discovery from fMRI images is proposed. In the voxel-centric approach, voxels are considered to be the units of activation – the signal characteristics are modeled for individual voxels, and activation levels for voxels are used as features for knowledge discovery. Here, an alternate *segment-centric* framework for knowledge discovery is presented – this framework is called Knowledge Discovery from Segmented fMRI (KDSf). In this framework, instead of characteristics of voxels, characteristics of functional segments are used for knowledge discovery. Functional MRI segments are clumps of voxels that are co-modulated by task-induced activation – these segments can be thought of as functional units (or functional ROIs [5, 6]) in the brain. The key assumption is that the MR signal patterns for functionally similar voxels behave in a similar fashion during the course of the experiment – this is commonly considered to be a valid assumption (e.g. see [7, 8]).

Functional units exhibit *spatial coherence*, which is a measure of homogeneity of behavior of physically adjacent voxels. The spatial coherence principle can be used to incorporate a number of voxels with similar temporal characteristics into a single functional unit (segment). Use of functional segments can potentially alleviate all three of the problems with voxel-centric knowledge discovery listed above.

The application of KDSf to image datasets requires a mechanism for functional segmentation of the 4D images, where, for each voxel location, a timecourse describes the evolution of the MR signal during the scan. The currently available methods for image segmentation are not suitable for unattended segmentation of timecourse-valued images. Auto-threshold Contrast Enhancing Iterative Clustering (ACEIC) – a new spatial coherence-based algorithm is presented here for this purpose [9].



This work is organized in three parts: first, it is determined if feature construction based upon functional segmentation (KDSf) is worthwhile – this is evaluated with simulated image datasets. Second, an algorithm (ACEIC) is presented for functional segmentation based on spatial coherence of time-courses. The accuracy of this algorithm is compared against Probabilistic Independent Component Analysis (ICA), a popular method used for similar purposes. Third, the spatial coherence principle is applied to feature selection for voxel-centric classification problems as well – this method, referred to as Spatially Coherent Voxels (SCV), is applied to a real-world classification problem.

### **1.2.1 Thesis**

First, it is hypothesized that feature construction via functional segmentation can improve classification accuracies from fMRI images when between-subject variability of regions of activation is not fully corrected by spatial normalization methods.

Second, it is hypothesized that the spatial coherence of timecourses can be exploited for functional segmentation of fMRI images.

Third, it is hypothesized that spatial coherence information can also be helpful for classification from un-segmented images (voxel-based classification).

Based upon the above theses, specific claims are made, two strong and another weak.

Strong claims:

1. Under controlled conditions (synthetic data), KDSf can achieve better classification accuracies than voxel-based classification, when spatial normalization is imprecise. For the purpose of this claim, it is assumed that some method of functional segmentation (with 100% accuracy) is available.
2. Under controlled conditions (public benchmark synthetic data), ACEIC demonstrates better accuracy of functional segmentation than Probabilistic Independent Component Analysis.

Weak claim:

1. Feature selection using spatial coherence information (SCV) can improve classification accuracy for voxel-based classification from real-world fMRI

datasets. This claim is tested for a classification problem of distinguishing between activation images based on a ‘neurobehavior disinhibition’ score.

### 1.3 SIGNIFICANCE

While the potential of fMRI is generally appreciated, the observed inter-subject variability of loci of activation has raised difficulties in assessing the clinical usefulness of fMRI. This inter-subject variability, along with the imprecise nature of inter-subject spatial normalization, presents a challenging problem for current methods of voxel-based analysis and knowledge discovery. In this work it is demonstrated that in the presence of inter-subject variability of loci of activation, the KDSf framework can provide better classification accuracies than the voxel-based approach. While this is demonstrated under controlled conditions only (with synthetic data), duplication of these results for real-world datasets will improve the appeal of fMRI for clinical applications.

Functional segmentation is a general data-reduction technique that can be applied for purposes other than machine learning. For example, functional segmentation with the ACEIC algorithm can be a first step for isolation of signal artifacts due to head-motion of the subject during an fMRI experiment. Also, the ACEIC method can be employed for exploratory analysis [10] and visualization of 4D data from sources other than fMRI experiments.

While application of the KDSf framework to real-world datasets would require improvements in functional segmentation techniques (extant methods are not suitable for *unsupervised* functional segmentation, and the ACEIC method is sensitive to head-motion as well as activation), the spatial coherence principle does also apply to voxel-based classification. This is demonstrated by the SCV method, which can be applied to real-world fMRI datasets without functional segmentation.

## 1.4 DISSERTATION OVERVIEW

Chapter 2 provides background information on fMRI technologies, conventional analysis pipeline for fMRI data, machine learning methods used in this work, methods of feature construction from maps of neural activation, and techniques for functional segmentation. For readers unfamiliar with fMRI, Chapter 3 provides a quick overview of a real-world fMRI experiment and the nature of discoveries made possible by fMRI. Chapter 4 describes the KDSf framework in detail and assesses the usefulness of the KDSf framework with synthetic data. Chapter 5 provides details about the ACEIC method for functional segmentation. Also, for a publicly available benchmark dataset, the segmentation accuracy of the ACEIC method is compared with that of Probabilistic Independent Component Analysis. Chapter 6 presents the SCV method of feature selection for voxel-based classification problems. For a real-world study of Substance Use Disorder (SUD), the accuracy of the SCV approach is compared with other methods of feature selection. Chapter 7 presents conclusions and discussions on further development of the methods presented here.

## **2.0 BACKGROUND**

### **2.1 IMAGING BRAIN FUNCTION**

While non-invasive ‘structural’ imaging of the brain has been possible since the advent of Computed Tomography (CT), ‘functional’ imaging of the brain is a relatively recent phenomenon. In functional imaging, the goal is to image the brain in action – this can localize cognitive function (or dysfunction) as opposed to lesions and other physical abnormalities observable with structural imaging. In CT, differential attenuation rates of x-rays can be used to produce detailed anatomical images of the brain. However, localization of functional activity in the brain requires some extrinsic or intrinsic contrast agent that can produce differential image intensity (contrast) at locations of neural activation. Emission tomography techniques such as Positron Emission Tomography (PET) and Single Photon Emission Computed Tomography (SPECT) use radioactively labeled contrast agents to identify regions of the brain exhibiting differential consumption of metabolites (typically glucose). However, the limited half-lives of these radioactive contrast agents, and the effects of ionizing radiation on the health of the subjects are some of drawbacks of the emission tomography methods. The development of the Functional Magnetic Resonance Imaging (fMRI) technique, which does not require the use of extrinsic contrast agents and is free of ionizing radiation, has provided a convenient platform for functional brain imaging for research purposes. The relative safety and the improvements in spatial and temporal resolution have made fMRI the imaging modality of choice for functional brain imaging studies.

Other techniques for observing brain activity include Magnetoencephalography (MEG) and Electroencephalography (EEG). In MEG, superconducting magnetometers are used to detect changes in magnetic fields due to electrical activity inside the brain – while this method has good temporal resolution, spatial localization of electrical activity from observed changes in magnetic

fields is a non-unique inversion problem. In EEG, surface electrodes are used to measure changes in electrical potentials due to brain activity – again, while the temporal resolution is better than fMRI, spatial localization of the activity inside the brain is difficult.

Optical imaging is an invasive imaging technique for monitoring brain activation by detecting activation-induced changes in absorption and scattering of light by exposed neural tissue. While, this method has good temporal and spatial resolution, the invasive nature of the method makes it unsuitable for human studies. Non-invasive optical tomography has also been used for imaging cortical hemoglobin oxygenation in humans, particularly for neonates.

### 2.1.1 Magnetic Resonance Phenomenon

The nuclear magnetic resonance (NMR) phenomenon was independently discovered by Bloch [11] and Purcell [12] around 1946. This phenomenon has been exploited for study of molecular structures and for detection of metabolites (spectroscopy). Lauterbur [13] first demonstrated the feasibility of construction of 2D and 3D images using the NMR phenomenon. Since then the development of Fourier reconstruction methods [14] has led to major advances in Magnetic Resonance Imaging (MRI).

Atomic nuclei with an odd number of protons and neutrons possess a property called spin which can be visualized as a rotation of the nucleus about its own axis. This spinning of charged nuclei creates a magnetic moment whose characteristics differ between types of nuclei (determined by the magnetic spin quantum number). The MRI technique exploits the behavior of protons (hydrogen atoms or ‘spins’) in a magnetic field. When a collection of spins are placed in a strong magnetic field, net magnetization of the collection aligns with the direction of the magnetic field (called the longitudinal direction). If such aligned spins are perturbed by injection of new energy, the net magnetization is displaced from the direction of the longitudinal magnetic field and precesses around the longitudinal direction with a frequency (Larmor frequency) that is proportional to the strength of the applied magnetic field – the proportionality constant is called the gyro-magnetic ratio. For protons, the gyro-magnetic ratio is 4257 Hertz/Gauss (the strength of the earth's magnetic field is approximately one Gauss). Thus at magnetic field strength of 1.5 Tesla (15,000 Gauss), the Larmor frequency for protons is 63.855 Megahertz.

$$\omega_0 = \gamma B_0 \quad (2.1.1)$$

where,  $\omega_0$  is the Larmor frequency,  $\gamma$  is the gyro-magnetic ratio and  $B_0$  is the strength of the longitudinal magnetic field. Gyro-magnetic ratios for some nuclei are listed in Table 1.

**Table 1.** Gyro-magnetic ratio for some nuclei [15].

<b>Nucleus</b>	<b><math>\gamma</math>(MHz/Tesla)</b>
$^1\text{H}$	42.58
$^{19}\text{F}$	40.08
$^{23}\text{Na}$	11.27
$^{31}\text{P}$	17.25

If a radio-frequency (RF) pulse matching the Larmor frequency (i.e. at the resonance frequency) is applied to a collection of precessing spins, the net magnetization direction of these spins is displaced from the longitudinal direction – the angle of displacement from the longitudinal direction is called the ‘flip’ angle (Figure 1). This introduces a component of the net magnetization vector into the plane perpendicular to the direction of the magnetic field (called the transverse plane). At the end of the RF pulse, the net magnetization vector gradually recovers its alignment with the external magnetic field as it continues to precess around the longitudinal direction with gradual decay of the angle of displacement from the longitudinal direction. This leads to a gradual decay of transverse magnetization along with a gradual recovery of the longitudinal magnetization. Detection coils placed in the transverse plane can detect this oscillating magnetic field and the signal (called free induction decay) can be interpreted for image creation. For image creation, the frequencies of the flipped spins are manipulated by varying the magnetic field strengths over space and time. Since functional scans require repeated imaging of the brain over time, the flipping process is repeated after the spins regain alignment with the longitudinal magnetic field.

The rate of recovery of the net magnetization  $M$  is called relaxation. Two factors influence the rate of relaxation – inhomogeneity of the applied magnetic field and the characteristic relaxation times for the tissues containing the spins. The relaxation time for a particular tissue type is characterized by two parameters –  $T_1$  and  $T_2$ . The dissipation of heat energy to the surrounding environment (lattice) is characterized by the  $T_1$  time, which is the time taken for approximately 63% of the longitudinal magnetization to be restored following a 90 degree pulse of RF energy.

$$M_z = M_0(1 - e^{-t/T1}) \quad (2.1.2)$$

where,  $M_z$  is the recovering longitudinal magnetization over time ( $t$ ) and  $M_0$  is the longitudinal magnetization at equilibrium.

T2 relaxation is the decrease in the transverse component of magnetization due to interaction between neighboring spins.

$$M_{xy} = M_{xy_0} e^{-t/T2} \quad (2.1.3)$$

where,  $M_{xy}$  is the transverse magnetization, and  $M_{xy_0}$  is the initial transverse magnetization at  $t=0$ , at the end of the RF pulse. T2 is always smaller than or equal to T1.

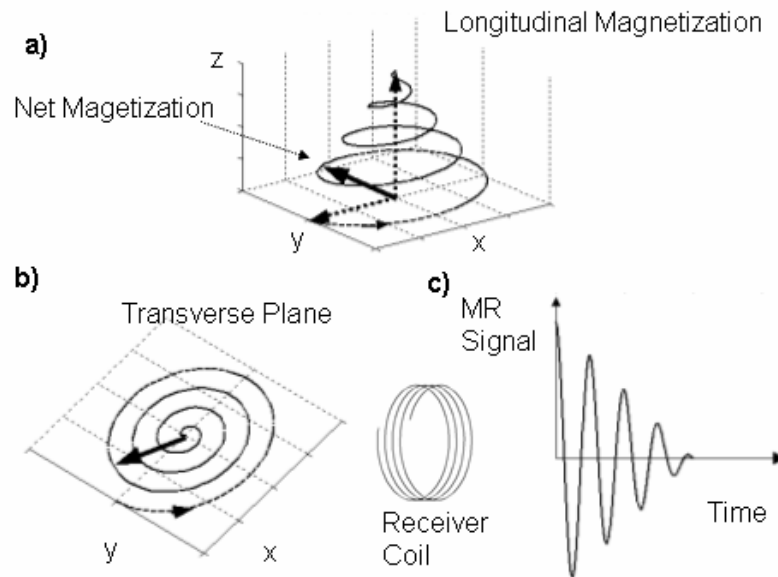
**Table 2.** Characteristic T1 and T2 parameters for different tissue types [15].

Tissue Type	T1 (1.5 Tesla) (milliseconds)	T2 (milliseconds)
Cerebrospinal Fluid	2400	160
White Matter	780	90
Gray Matter	900	100
Muscle	870	45
Adipose	260	80

However, the decrease in the transverse component is also affected by the inhomogeneity of the magnetic field which causes the spins to rotate with slightly different frequencies causing loss of coherence and signal decay. The decrease in transverse magnetization (which does not involve the emission of energy) is called decay. The rate of decay is described by a time constant, T2\* which is the time taken for the transverse magnetization to decay to 37% of its original magnitude. This decay of transverse magnetization (characterized by T2\*) includes effects of spin-spin interactions (T2) and effects of magnetic inhomogeneity – T2\* is always smaller than T2.

The transverse magnetization signal decays as the precessing spins lose coherence (dephase) over time – magnetic gradients can be used to ‘rephase’ the spins to generate a new signal called an ‘echo’. The delay between the original RF excitation pulse and the signal echo from the spins is called TE (echo-time). The MR signal is recorded (read out) starting from TE. For a given TE, tissues with longer T2 have higher signal intensity than tissues with shorter T2 (quicker decay). The interval between applications of the RF excitation pulses (after recovery of

longitudinal magnetization) is called TR (repeat-time). For a given TR, tissues with short T1 have greater signal intensity than tissues with a longer T1 (slower recovery of longitudinal magnetization). Thus, TR and TE can be adjusted to control the differences in signal intensities (contrast) from different tissues types with different T1 and T2 times. Human brain has T1 around a second (with higher values for CSF) and T2 of about 100 milliseconds at typical imaging field strengths (Table 2).



**Figure 1.** Nuclear Magnetic Resonance. a) Precession of the net magnetization vector around the longitudinal magnetic field while recovering from a 90 degree flip to the transverse plane (caused by introduction of an RF pulse). b) The receiver coil in the transverse plane detects the signal from the precessing magnetization vector. c) Decaying echo signal recorded (read out) from the receiver coil.

### 2.1.2 MRI Pulse Sequences

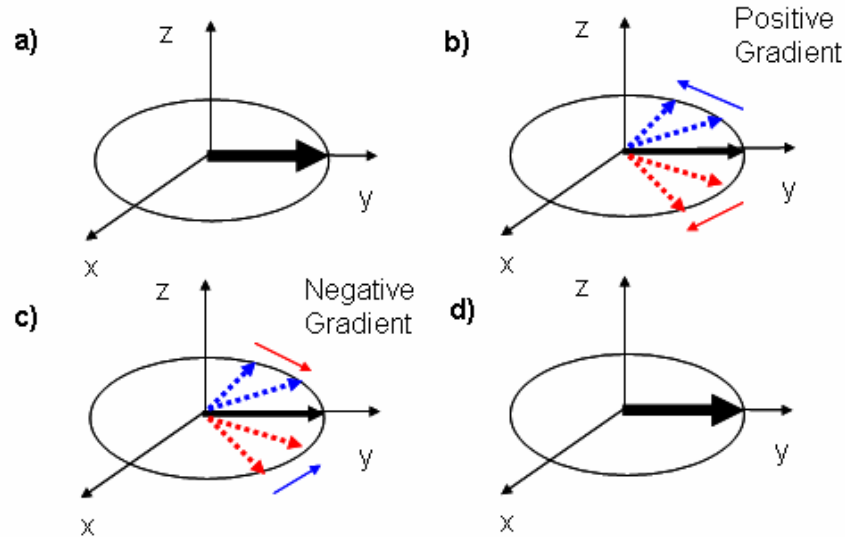
Since magnetic inhomogeneity causes spins to rotate with slightly different Larmor frequencies, this can cause dephasing or loss of coherence between the spins leading to loss of signal intensity. Special techniques are used to refocus the spins to maximize the signal strength.

Sequences of RF pulses called ‘spin echo pulses’ can be used to resynchronize the spins so that the decay rate is controlled by the T2 parameter rather than T2\*. A 180 degree pulse applied after the original 90 degree pulse reverses the developing phase relationships between



spins due to magnetic inhomogeneity – this leads to resynchronization of the spins and generation of a new signal echo.

Alternatively, to obtain images where the contrast behavior incorporates contributions from  $T2^*$ , a ‘gradient echo’ contrast technique can be used. Here, instead of a 180 degree RF pulse, reversal of magnetic gradient is used to synchronize the spins (Figure 2). Also, flip-angles smaller than 90 degree used with gradient echo images are suitable for fast imaging.



**Figure 2.** Gradient Echo technique. a) All protons are in phase. b) When gradient is applied, some protons spin faster than before (red) and some slow down (blue), leading to loss of signal. c) Reversing the gradient speeds up the previously slowed down protons and vice versa. d) Protons are in phase again and produce a new signal echo.

A pulse sequence is an appropriate combination of one or more RF pulses and magnetic field gradients with intervening periods of recovery of longitudinal magnetization. A pulse sequence is characterized by several parameters, including the repetition time (TR) between repeated excitations of the same spins, the echo time (TE), and the flip angle. The application of RF pulses at different TRs and the receiving of signals at different TEs controls ‘weighting’ of the images to emphasize different tissue-types – fMRI is typically  $T2^*$ -weighted since the differences in blood oxygenation levels leads to different  $T2^*$  times for the neural tissue.

### 2.1.3 Spatial Encoding

The signal received by the receiving coil is the sum of the signal from all the spins in the object being scanned (the brain for fMRI). The creation of an image requires characterization of the

spins at different spatial locations. This is achieved by applying magnetic gradients to assign different frequencies (and phases) to spins at different locations in the brain. Since frequencies are used to encode spatial information about the spins, recovery of the spatial information requires a Fourier transformation of the signal recorded from the receiver coil. The time-varying magnetic gradients can be thought of as traversing the space of the (spatial) Fourier transform of the object in the scanner (called k-space or spatial-frequency domain). Alternatively, k-space is a representation of the MRI raw data before it has been Fourier-transformed in order to make an image. The signal location in k-space is the integral of the magnetic gradient amplitude and ‘on-time’ of the gradient – at any point of time  $t$ , the phase of the signal from a voxel is determined by the cumulative effects of the rotational speed-up caused by gradients applied till that point in time. The MR signal equation for the whole volume is

$$S(t) = \int_V M_0(\vec{r}) \sin(\alpha) e^{i\omega_0 t} e^{i2\pi \vec{K}(t) \cdot \vec{r}} d^3\vec{r} \quad (2.1.4)$$

where

$\vec{r}$  is the location of a voxel in 3D,

$V$  is the volume in the scanner,

$M_0(\vec{r})$  is the equilibrium magnetization at location  $\vec{r}$ ,

$\alpha$  is the flip angle,

$\omega_0$  is the Larmor frequency at  $B_0$ , the longitudinal magnetic field strength,

$\vec{G}$  is the time-varying magnetic gradient vector (3D),

and  $\vec{K}(t) \cdot \vec{r}$  is the cumulative speed-up effect (up to time  $t$ ) of the gradient at location  $\vec{r}$

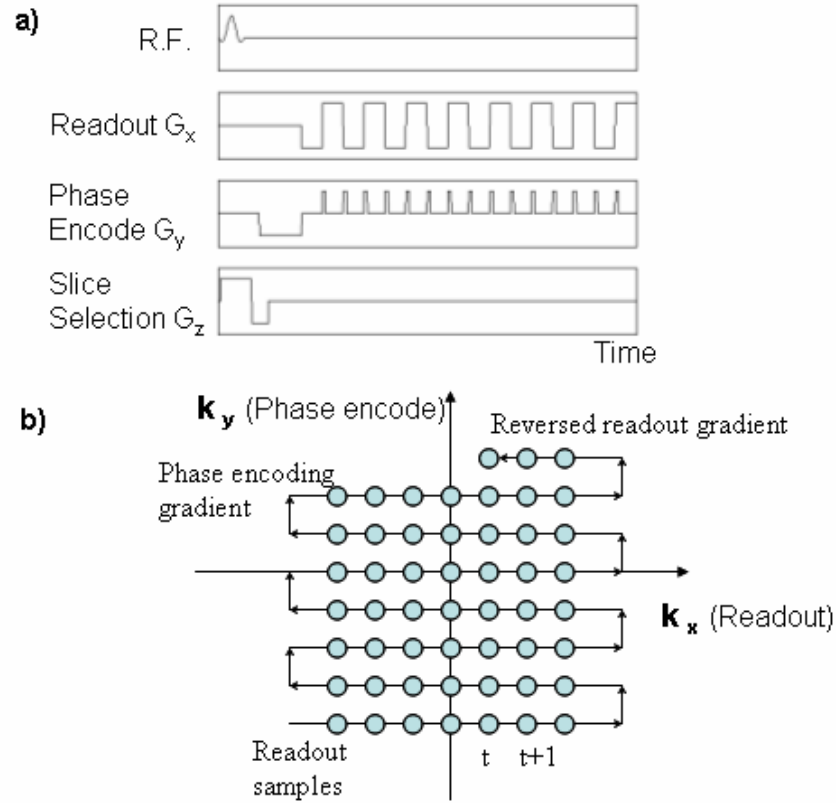
$$\vec{K}(t) \cdot \vec{r} = \frac{\gamma}{2\pi} \int_0^t \vec{G}(s) \cdot \vec{r} ds \quad (2.1.5)$$

The form of the signal equation shows that  $\vec{K}$  is the conjugate spatial-frequency domain corresponding to the spatial domain  $\vec{r}$ , and the readout signal  $S(t)$  covers k-space  $S(\vec{K}(t))$ .

This traversal of k-space can take various trajectories – in spiral imaging, a spiral trajectory through k-space is used; in echo planar imaging (EPI), a back-and-forth motion covers k-space of a 2D slice [16, 17]. In 2D imaging (EPI), a longitudinal gradient is used to select the

image slice and two perpendicular directions (called frequency encoding direction and phase encoding direction) are used for spatial encoding within the slice. The RF pulse is synchronized with the longitudinal gradient so that only the spins in the slice are excited ('flipped') by the RF pulse. In 'single-shot' EPI, the whole slice is imaged from a single excitation of the spins – typical imaging time is around 60 milliseconds per slice (for example, 24 slices can be imaged within  $TR=1.5$  seconds, with  $TE=30$  milliseconds). The gradient applied in the phase encoding direction for a short duration speeds up the spins yielding a phase difference between the spins along that direction (called phase memory). The gradient applied in the frequency encoding direction (or readout direction) is turned on during signal recording to frequency-encode the spatial location of the spins in the readout direction. Thus each sample from the 'quadrature' receiving coil is a complex-valued entity (with phase information) that populates the k-space of the image. After the k-space is sampled, Fourier transform of the k-space data provides an image of object (this step is called image reconstruction). A schematic of the EPI pulse sequence and the corresponding k-space trajectory (as the signal from the receiver coil is sampled over time) is shown in Figure 3.

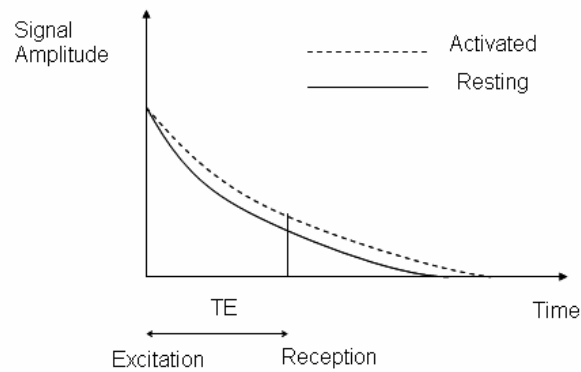
For repeated (functional) imaging of the brain volume, the spins in a slice are repeatedly excited after recovery of the longitudinal magnetization. In 'single-shot' EPI (see above), where the whole slice is imaged from a single excitation of the spins,  $TR$  is also the time required to acquire one image volume (the slices are excited independently). Thus, in this situation, the fMRI signal for a voxel is sampled at  $TR$  intervals.



**Figure 3.** Echo Planar Imaging (EPI) a) Pulse sequence for Echo Planar Imaging. b) k-space trajectory as the signal from the receiver coil is sampled over time.

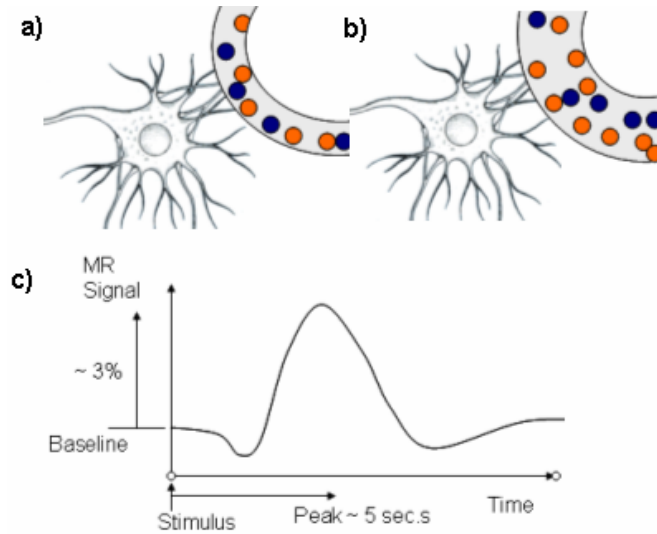
#### 2.1.4 BOLD Contrast

Changes in blood oxygenation level can be observed with MRI. Oxy-hemoglobin (oxygenated state of the oxygen-transporting protein hemoglobin) has no substantial magnetic properties, but deoxy-hemoglobin (after oxygen is delivered to the tissues) is strongly paramagnetic. Thus, deoxy-hemoglobin ions create oscillating magnetic fields in their neighborhoods resulting in positive magnetic susceptibility. The presence of these oscillating magnetic fields increases the rate of decay of the transverse magnetization from precessing spins ( $T2^*$  is reduced). Thus, the increase in the proportion of oxy-hemoglobin in the blood (due to over-compensation for increased oxygen consumption) increases the  $T2^*$  parameter for the local spins, which leads to an increase of signal intensity (Figure 4). Hence, deoxy-hemoglobin can serve as an intrinsic paramagnetic contrast agent that is sensitive to neural activation.



**Figure 4.** Decay of transverse magnetization is slower during activation (larger  $T2^*$ ) – thus, MR signal amplitude is higher for activated state.

In the brain, oxygen is passively transported from oxy-hemoglobin to the plasma, then to extra-vascular space (interstitial space), to the intra-cellular space, and finally reaches the mitochondria via a pressure gradient – the concentration of oxy-hemoglobin in blood maintains this pressure gradient [18]. During neural activity, blood oxygenation levels in the capillaries and venules surrounding the neural tissue exhibit an early drop followed by a gradual rise. The oxygenation level achieves a plateau with continue neural activity. On cessation of activity, the oxygenation level returns to the baseline, and may eventually undershoot it. These neuro-vascular couplings produce a complex MR signal function corresponding to brief neural activity – this is referred to as the hemodynamic response function (HRF) and is shown in Figure 5.



**Figure 5.** Hemodynamic response to brief stimulus. a) Resting state. b) Neural activation is associated with higher concentration of oxy-hemoglobin (red circles) in the blood compared to the resting state. c) Idealized hemodynamic response function.

During activation, the oxygen consumption by the local neural tissue increases by approximately 5% leading to localized dilation of blood vessels and a local increase of blood volume and flow by 20 - 40%. The elevated supply of oxygenated blood reduces the ratio of deoxy-hemoglobin to oxy-hemoglobin, compared to the basal state. The resulting change in  $T2^*$  can be detected by an appropriately designed pulse sequence. For example, by using  $T2^*$ -weighted gradient echo EPI sequences, a 2-5% increase in signal intensity (proportional to the underlining neural activity) can be observed 4-6 seconds after onset of neural activation (Figure 5). Note that the inherent noise level of the signal is also of the order of 2-3%. The ratio of activation-induced rise in the signal level to the noise level in the signal is called the contrast-to-noise ratio (CNR). Also, since neural activation is indirectly inferred from hemodynamic changes (local increases in blood flow and oxygenation levels), spatial localization of neural activation is only approximate – spatial resolution is considered to be around  $3 \text{ mm}^3$ , which may correspond to half-million neurons in the human cortex. Similarly, temporal resolution is of the order of several seconds.

During the course of the fMRI experiment (typically lasting 5-10 minutes), a time-series of image volumes is acquired continuously (one image volume every 1-3 seconds). The fluctuations in blood oxygenation are observed as spatially localized temporal fluctuations in signal intensity in the re-constructed series of images. Thus, information about dynamics of

neural activation during the experiment are delayed and blurred by the hemodynamic response to neural activation.

### **2.1.5 Limitations of fMRI data**

Interpretation of the fMRI images is complicated by the noisy nature of fMRI images, caused by technical factors (e.g. thermal noise from electronics, scanner instabilities etc.) and artifacts introduced by physiological factors (e.g. inadvertent movements of the subject's head, cardiac pulses etc.) [15].

#### **2.1.5.1 Technological Limitations**

The signal-to-noise ratio (SNR) of individual Magnetic Resonance images is impacted by various factors. The SNR has an approximately linear dependence on the strength of the external magnetic field – higher fields generate a stronger net magnetization vector (and its transverse component) leading to stronger signal in the coil. However, for very high field strengths, the improvement in SNR is offset by increased T1 times and generation of artifacts. The SNR is also dependent on the number of spins contributing to a voxel value – thus, coarser spatial resolution improves SNR. For fMRI, the particular choice of scanning parameters achieves a trade-off between SNR, resolution and time needed for acquisition of individual volumes.

Artifacts are characteristics of images that do not reflect actual tissue/phenomenon being imaged. Factors contributing to MRI artifacts include 1) inhomogeneity of the external magnetic field, 2) magnetic susceptibility artifacts caused by local field variations at air-tissue interfaces in the head, and 3) improper choice of imaging parameters (e.g. wrap-around aliasing, Gibbs ringing etc.). The susceptibility artifact causes the orbito-frontal cortex and the anterior temporal lobes to have very low signal-to-noise ratio. Thus, the absence of BOLD activity in these regions should be interpreted cautiously.

#### **2.1.5.2 Physiological Limitations**

It is not possible to eliminate head-movement completely – even with co-operative subjects and the use of head-restraints. Tasks requiring overt speech responses pose particular difficulties with head-motion [19]. The signal changes introduced by head-motion of the subject

can be correlated with the stimulus, leading to detection of spurious activations. The amplitude of signal changes due to head-motion may exceed that of neural activation. Thus, motion-correction steps are employed to attempt to remove the motion-induced effects prior to data analysis – however, as demonstrated in Chapter 5, some effects may survive the application of motion-correction procedures.

Cardiac and respiratory cycles are the major sources of physiological noise in fMRI data. The cardiac effects include pulse-related local motion and hemodynamic changes. For typical image acquisition rates (e.g. one image volume every 2-3 seconds), the cardiac signal is aliased, making it difficult to correct for the cardiac signal without concurrent recording of cardiac events. Respiratory artifacts may be induced by changes in magnetic field homogeneity due to moving organs, respiration-dependent vasodilation, or changes in oxygenation levels. There is some evidence that the respiratory effects are localized in white matter only [20].

While PET can provide relatively accurate measurements of blood flow and other metabolic measures, fMRI signal values cannot be interpreted quantitatively – it is necessary to compare the signal values with a reference state. The BOLD response is an indirect measurement of neural metabolism – blurred and delayed by the hemodynamic response function.

## **2.2 CONVENTIONAL DATA ANALYSIS PIPELINE**

Due to the low signal-to-noise ratio of fMRI data, the robustness of the results depends upon the choice of experimental protocol and data analysis method. These choices include 1) the experimental ‘paradigm’, 2) data preparation steps prior to statistical analysis, 3) statistical analysis of data from individual subjects, and 4) pooling of evidence from different subjects. The design of the experimental paradigm is based upon the cognitive process of interest and limitations of the scanner environment (space and noise issues). The data preparation steps compensate for the noisy nature of the fMRI signal and the artifacts introduced by inadvertent head-motion of the subject. This is followed by statistical analysis to yield a brain-map of neural activation in individual subjects. Finally, the evidence from individual activation maps is pooled to yield a consensus brain-map of the cognitive process of interest. Statistical or machine



learning techniques can be employed for discovery of differences in activation patterns between groups of subjects that may be useful for clinical applications.

## **2.2.1 Experimental Design**

### **2.2.1.1 Experimental Paradigm**

Paradigm is defined as the construction, temporal structure, and behavioral predictions of cognitive tasks executed by the subject during an fMRI experiment [18]. The paradigm is designed to isolate some brain process of interest – e.g. a task requiring inhibition of reflexive eye-movements may be employed to isolate cortical regions involved in inhibitory processes. Since it is not possible to measure absolute values of neural activation, a basis for comparison needs to be chosen. The subtraction principle assumes that the cognitive components are additive and there are no interactions between cognitive components of the task. Though this is invalid in general, this has proven to be useful for fMRI studies. Using this principle, any statistically significant difference in signal levels between two conditions is attributed to the cognitive process that differentiates the two conditions. For example, to isolate brain centers for language production, two conditions might be employed – one in which the subject vocally repeats a word presented on screen, and another during which the subject views the word without vocalization.

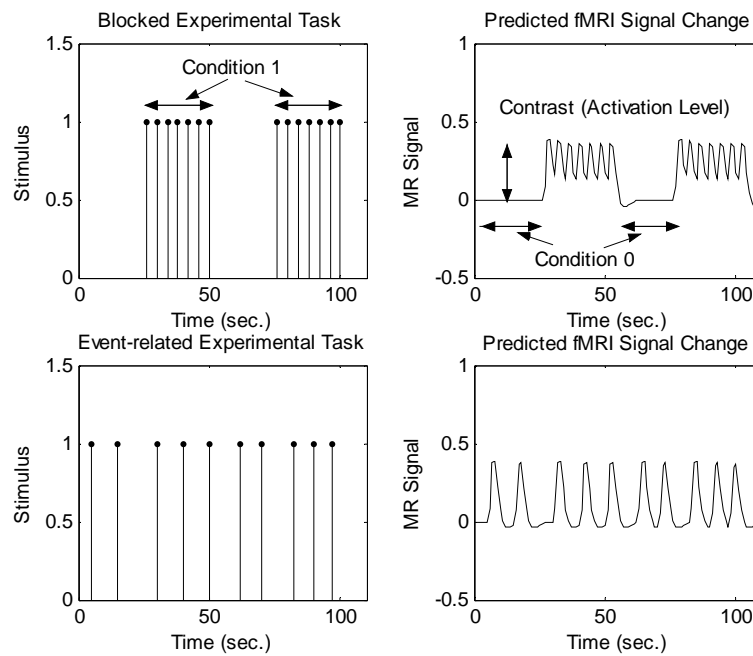
Interactions between cognitive processes can be studied with factorial designs [18] where conditions consist of different combinations of a few cognitive processes – this can also be used to test the validity of the subtraction principle. Since some tasks can involve different levels of difficulty (e.g. mental rotation of objects), parametric designs attempt to isolate regions of the brain that mirror the parametric manipulation. However, this can pose problems if different regions of the brain are recruited as task difficulty is manipulated.

### **2.2.1.2 Stimulus Presentation**

During a run of the experiment, the scanner continuously collects 3D image volumes yielding a signal time-series for each voxel location in the brain. Due to the noisy nature of the MRI signal and low Contrast to Noise ratio (CNR) of BOLD activation, measurements are averaged over multiple repetitions of the task. Two main strategies are employed for signal averaging – block designs and event-related designs. In block designs, the same task is performed during a

continuous block of time – for example, 30 seconds of language-production may be alternated with 30 seconds of rehearsal, then the whole cycle may be repeated a few times. Due to its higher statistical power and large differences in the saturated BOLD signal between two conditions, block designs generally yield more robust results [18].

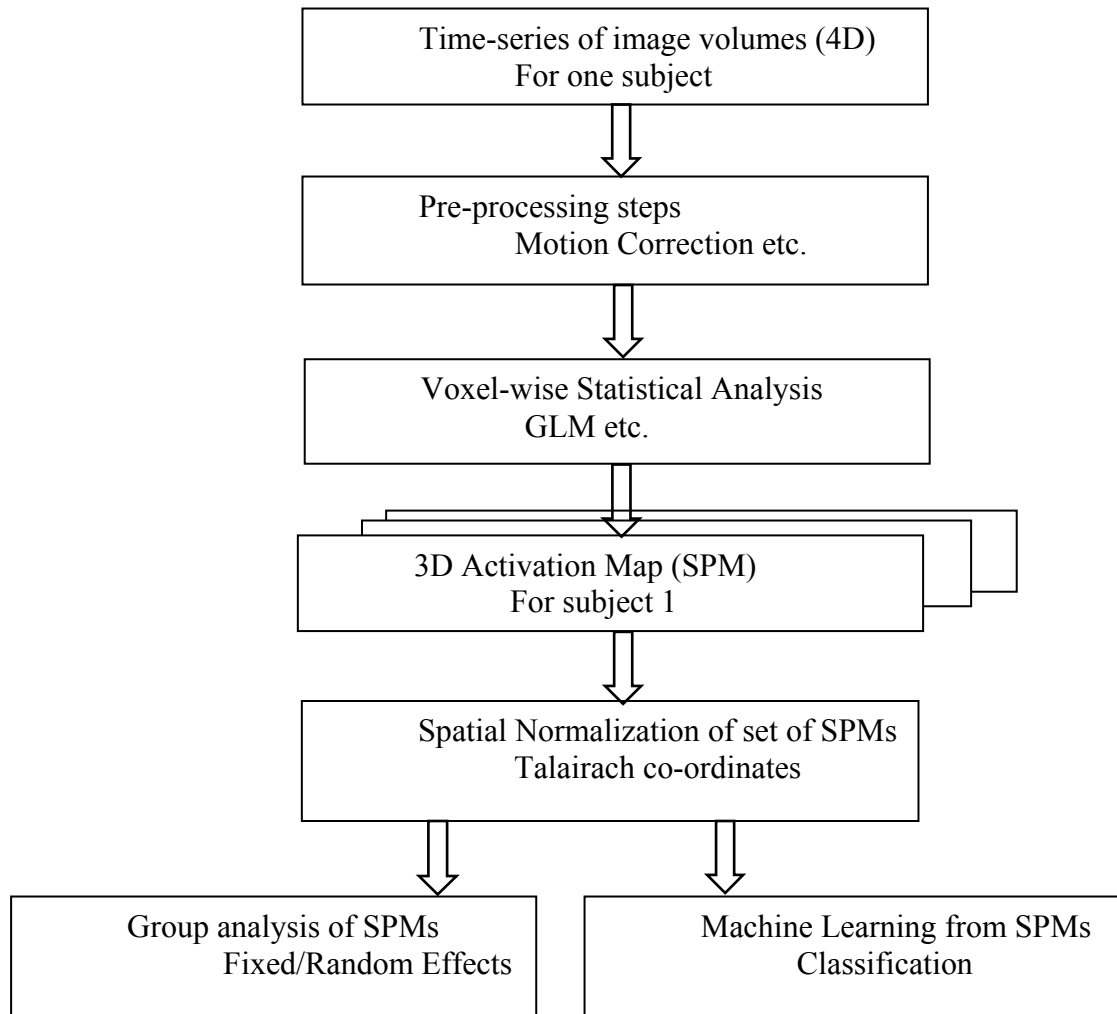
In event-related designs [19], tasks are individual trials of short duration and may be interspersed with other tasks – for example, a word of a particular semantic category may be presented to the subject at random intervals. The activation response is averaged over multiple trials of the same type and is compared with the basal signal level or with the activation response from another type of trial (e.g. a different semantic category). Event-related designs allow for characterization of the hemodynamic response for individual trials – the differences in the hemodynamic response at different locations in the brain can also be detected. This approach is also insensitive to head-motion artifacts and it allows the identification of neural correlates of behavioral responses (e.g. correct/incorrect responses to the trials). However, event-related designs can require longer scan times since the inter-stimulus interval (ISI) must be long enough for sufficient decay of hemodynamic response. Also, compared to the block design, this approach is more sensitive to assumptions about the shape of the HRF and linearity of overlapping HRFs. These two approaches to experimental design are illustrated in Figure 6.



**Figure 6.** Comparison of Block (top row) and Event-related (bottom row) task designs.

### 2.2.2 Analysis pipeline

The MR scanner collects data in the spatial frequency domain (k-space) – a subsequent reconstruction step is required to transform the k-space data to brain images. As the subject performs the tasks, the scanner continuously collects k-space data, sampling the designated volume of the brain every few seconds (typically 1-3 seconds). After the experiment, the sequence of image volumes is reconstructed from the sequence of k-space volumes by Fourier transformation. Thus, after reconstruction, the fMRI experimental data is a 4D spatio-temporal volume tracking MR signal fluctuations at different locations of the brain. A sequence of data-processing steps is applied to this 4D dataset to enhance the signal-to-noise ratio (SNR) and to eliminate artifacts introduced by inadvertent head-motion of the subject during the experiment. This is followed by statistical analysis of the 4D data to yield a 3D brain-map of activation. Before group comparisons, 3D activation maps from the set of subjects are transformed (‘spatially normalized’) to a common co-ordinate system. Finally, statistical analysis or machine learning methods are employed for knowledge discovery from this ‘normalized’ set of 3D activation maps. The steps in the pipeline are shown in Figure 7.



**Figure 7.** Conventional data-processing pipeline for fMRI data

### 2.2.2.1 Pre-processing steps

A sequence of pre-processing steps can be applied to the data prior to statistical analysis. These optional steps are designed to remove known sources of error prior to statistical analysis.

#### Slice-timing correction

Since each image volume is acquired slice by slice (for EPI), there is a lag between sampling times for different slices in the volume. Interpolation techniques are used to correct for this temporal misalignment which may manifest itself as latency differences in the hemodynamic response of different voxels.

#### Motion Correction

Head-motion of the subject during the experiment introduces various artifacts [21] to the fMRI time-series – typically these artifacts are distributed around the outer periphery of the brain images. The amplitude of these motion-related changes to the signal can be significantly higher than the BOLD response. If these artifacts are correlated with the task, they can confound the results of the statistical analysis.

To correct for the effects of head-motion, individual image volumes in the time-series are registered with a reference image volume (usually the first image volume or the mean of the image volumes) [22]. The registration problem is typically modeled as a search for optimal parameters of 3D rigid-body transformation (pitch, roll, yaw and translation) and the objective function is based on the voxel-wise differences in image intensities between the two images volumes being registered. For each volume in the series, the optimal transformation is determined and applied to the individual volume for realignment. Application of the transformation requires interpolation between voxels, which may in turn introduce interpolation artifacts [23].

#### Physiological Correction

Cardiac pulses and respiratory cycles can impact the fMRI time-series. If additional physiological data is collected during the fMRI experiment, corrections for these effects can be applied. The corrections can be applied in k-space [24] or in image-space [25].

#### Spatial Smoothing

Several motivations suggest spatial smoothing of the image volumes in the time-series prior to statistical analysis. First, spatially smoothing can improve the signal-to-noise ratio (SNR) of the activated regions. Typically, a 3D Gaussian smoothing kernel is convolved with the image volume. The Full-Width-at-Half-Maximum (FWHM) parameter of the smoothing kernel determines the degree of smoothing. The optimal choice of the kernel-width depends upon the spatial extent of activation in the image – over-smoothing can occur with kernels that are wider than the region of activation. Typically FWHM values between 4-8 mm are used for this purpose.

Also, the central limit theorem suggests that smoothing will cause the distribution of the errors in images to be more Gaussian-like, which will improve the validity of the single-voxel inferences from parametric tests. Also, statistical inferences at the level of clumps of voxels

based on Gaussian Random Fields also assume a lattice representation of a smooth Gaussian field.

Finally, smoothing can provide a measure of relief from inadequacies of inter-subject spatial normalization (described below). However, smoothing also introduces loss of spatial resolution.

#### Temporal Smoothing

As with spatial smoothing, the signal time-series can also be temporally smoothed to improve the signal to noise ratio – Gaussian kernels can be used for this purpose. Note that temporal smoothing introduces additional temporal dependence between signal values – this may invalidate the independence assumptions inherent in many statistical tests. Also, a high-pass filter can be used to remove low-frequency trends from the time-series.

#### **2.2.2.2 Voxel-wise Statistical Analysis**

For purposes of data analysis, the fMRI signal time-series for a voxel is typically modeled as a linear time-invariant system. The hemodynamic response to the MR signal from a brief (approximately 1 second) period of neural activity is treated as the impulse response function of a linear system [26]. This hemodynamic response function (for brief neuronal activity) is used to create a statistical model of the expected signal time-series, given the history of stimulus presentation (Figure 6). The conformance of the observed time-series with the expected time-series is assessed by regression techniques, yielding a statistical parameter value representing the observed degree of conformance (or activation strength). The voxels for which this conformance is statistically significant is labeled as ‘activated’ by the task.

For each voxel, independent statistical models are fitted to the voxel time-series and the representation of these statistical parameter values (scores from statistical tests) on a 3D map of the brain is called the Statistical Parameter Map (SPM) or Activation Map. Some common statistical models used for this purpose are described below.

#### Statistical Models

##### a) T-test

For simpler block-design experiments, a two-sample Student’s t-test can be used to assess the significance of differences in activation levels for two task conditions (assuming signal saturation during the block, after the onset delay in hemodynamic response). For example, the

signal during performance of the language-production task can be compared with the signal during language-rehearsal. The t-score compares the mean of the set of signal values for one condition ( $X_1$ ) with the mean of the set of signal values for the other condition ( $X_2$ ), relative to a pooled estimate of the variability of the difference.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} \quad (2.2.1)$$

where, the standard deviation of the difference of the means is given by

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \quad (2.2.2)$$

and the pooled variance  $s_p^2$  is computed by

$$s_p^2 = \frac{\sum (X_1 - \bar{X}_1)^2 + \sum (X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2} \quad (2.2.3)$$

#### b) General Linear Model (GLM)

The two-sample t-test described above is a special case of the General Linear Model [27] which tries to explain a series of observations for a voxel in terms of explanatory conditions such as task performance. The GLM is formulated as

$$Y = X\beta + \varepsilon \quad (2.2.4)$$

where

$Y$  is a vector of time-point signal observations for a voxel, and

$X$  is the ‘design matrix’ where each column specifies the state of each of the explanatory variables for the time-points of observation, and

$\beta$  is a vector denoting the contributions of explanatory variables (e.g. activation strengths corresponding to a particular *conditions*), and

$\varepsilon$  is the residual error, assumed to be Gaussian with zero mean.

Each column of the design matrix specifies the expected response of the signal corresponding to a particular explanatory variable, given the stimulus presentation history and the assumed shape of the hemodynamic response (Figure 6). Inclusion of other explanatory

patterns (e.g. temporal derivatives of the explanatory variables) in the design matrix can also improve the fit.

The least-squares estimate of the explanatory variables is given by

$$\hat{\beta} = (X^T X)^{-1} (X^T Y) \quad (2.2.5)$$

Differences in activation levels between conditions (called contrasts) are calculated as differences between elements of  $\hat{\beta}$ . Different comparisons of interest are represented with different contrast vectors  $\lambda$ . For example, a simple contrast vector  $[-1 \ 1]^T$  can be used to compare two different types of trials (conditions). The contrast t-statistic is computed as

$$t_{\lambda} = \frac{\lambda^T \hat{\beta}}{\sqrt{\text{Var}(\lambda^T \hat{\beta})}} \quad (2.2.6)$$

This statistic is used for voxel-level assessment of statistical significance of the observed contrast, against the null hypothesis of no difference between conditions of interest. (Details about the computation of  $\sqrt{\text{Var}(\lambda^T \hat{\beta})}$  can be found in [28].) As with the t-test, this t-score can be mapped to the location of the voxel in the brain to yield a 3D statistical parameter map (SPM). An example SPM (activation image) is shown in Figure 17.

### Statistical Inference

Statistical inference can be employed for labeling a set of voxels as ‘significantly activated’ by the task of interest. The risk of falsely labeling a voxel as activated can be controlled at the voxel level by thresholding the t-scores (or z-scores) at the value corresponding to the chosen risk level (critical p-value). However, due to the large number of voxels under consideration, the threshold should be adjusted for multiple comparisons. Since hundreds of thousands of voxels are tested for significance, standard methods of correction for multiple comparisons (e.g. Bonferroni correction) are too stringent for in this situation. An alternative thresholding method employs Gaussian Random Field theory to assess the significance level of a clump of  $k$  voxels with z-scores higher than a threshold  $u$ . While this approach takes into account both the spatial extent and the strength of the activation, it is sensitive to the assumptions in the model. Other possibilities include the use of non-parametric permutation tests, where the inferences are free of model assumptions, at the expense of additional computational cost.



### 2.2.2.3 Spatial Normalization

Since locations of brain activation vary between subjects, multiple subjects are studied with the same experimental protocol to obtain a consensus pattern of activation in the population. However, since there are morphological differences between individual brains, the brains images need to be normalized to a common co-ordinate system before generalization to a population. This is done by warping the individual image volume into registration with a template that conforms to an anatomical atlas [29]. This atlas uses a proportional grid system based on the location of two anatomical landmarks – the Anterior Commissure (AC) and the Posterior Commissure (PC).

Spatial normalization methods can be broadly grouped into methods that employ anatomical labels and those that do not employ anatomical knowledge. The labels are usually manually identified based on landmarks and features of the images – the goal of the image transformation method is to superpose the labeled points. Since the labeling process is subjective, reproducibility of results can vary. Alternatively, instead of alignment of labels, some index of the difference between the two images can be minimized – typically, intensity differences between voxels are considered. The transformations are constrained to be slowly varying over space – thus contiguous voxels tend to remain contiguous after transformation. Several methods are available for spatial normalization: 12-parameter affine transformations, convolution-based transformations (with spatial basis functions), high-dimensional vector fields providing mappings for each voxel, and viscous fluid models [30].

The required transformation parameters for a subject are computed by registration of the structural image volumes of the individual (T1-weighted, acquired at the time of the experiment) with the template image. Once the required parameters are computed, the transformation can be applied to the time-series of functional volumes prior to statistical analysis, or to the final activation map (SPM) of the subject after statistical analysis for the individual. Spatial normalization can be imprecise – automated normalization tools may not be able to achieve voxel-level correspondence between images for a set of subjects. For example, Nieto-Castanon et al. [5] used anatomical markers to manually identify 10 ROIs from perisylvian cortical areas in the temporal and parietal lobes and compared the voxel-overlap between these ROIs across 9 subjects after the T1-weighted images were spatially normalized with the Matlab-based SPM program [31]. In this study, one of the ROIs (right hemisphere Heschl's gyrus) showed a mean

overlap of 31.43% across two normalized subjects, 13.26% across three normalized subjects and no overlap across all nine normalized subjects. The results for the other ROIs were similar – only the largest of the 10 ROIs showed non-zero (less than 5%) overlap across all nine subjects.

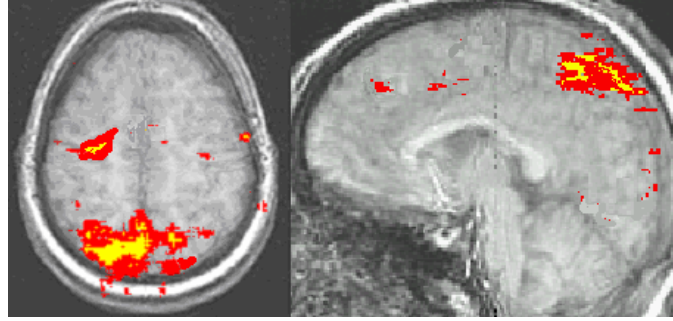
While spatial smoothing may bring some of the voxels into registration, it also causes loss of spatial resolution – there is no consensus regarding the optimal degree of smoothing. Smoothing can also impact the ability to detect between-group differences in activation levels (see Chapter 4). This work proposes segment-centric machine learning to address this problem with spatial normalization.

#### 2.2.2.4 Group map from multiple subjects

Once all the activation maps (SPMs) from different subjects are morphed to a common anatomical template by spatial normalization, the normalized activation maps can be compared across subjects to pool the evidence for task-induced activation at each voxel location. Group maps of brain activation can be created using a variety of methods for combining statistical information [32]. In the following it is assumed that for each subject and for each voxel location, p-values are computed from the statistical test used to detect task-induced activation. If  $k$  subjects are part of the study, the p-values  $P_1, P_2, \dots, P_k$  for each voxel location can be pooled to provide a measure of the cumulative evidence for activation at the particular voxel location. The Stouffer method combines the p-maps created from analysis of individual brains to create a combined group t-map. This method defines a combined test statistic

$$T_S = \sum_{i=1}^k \frac{\Phi^{-1}(1 - P_i)}{\sqrt{k}} \quad (2.2.7)$$

where  $\Phi^{-1}$  is the inverse normal cumulative distribution function. Combining statistical data with this method allows for robust group results while minimizing the effects of individual outliers [33]. The null hypothesis of no significant activation at the voxel location is rejected for large values of the test statistic, determined from critical points of the standard normal distribution. Again, due to the large number of voxels, voxel-wise statistical inference should be adjusted for multiple comparisons. An example group activation map is shown in Figure 8.



**Figure 8.** Example group activation map (thresholded by t-score and overlaid on anatomical image of template brain).

### 2.2.2.5 Comparison between groups of subjects

To compare the activation patterns between groups of subjects, the group activation maps from the two groups can be compared visually. Alternatively, voxel-wise fixed effects or random effects analysis [34] can be performed to identify locations of significant between-group differences in activation levels. Note that using the voxel-based significance-testing approach, it is not possible to directly test hypotheses regarding differences in sizes (spatial extents) of activated regions between groups. While larger activated regions in one group of subjects may lead to higher between-group t-scores around the periphery of the consensus region of activation, this effect can be unreliable in the presence of inadequate normalization of inter-subject variation in the location of activation. The KDSf framework addresses this particular problem.

## 2.3 MACHINE LEARNING METHODS

Creation of computer-based models of patterns in data is referred to as Statistical Learning or Machine Learning – similar activities are also referred to as Pattern Recognition, and Knowledge Discovery and Data-mining (KDD). The advent of the digital age has created large depositories of information which are a potential source of unexpected insights – the automated search for these insights is broadly referred to as data-mining or automated knowledge discovery. Machine learning or statistical learning refers to the creation of computer-based models from data that can be used for predicting something useful from data.

Since machine learning methods are designed to predict the value of some variable in terms of other variables, this does not directly lead to knowledge discovery – the machine learning problem is a ‘fitting’ problem where a statistical model is fitted to the data. In conventional machine learning, the learning model, the predictor variables, and the predicted (outcome) variable are all specified by the researcher. For automated knowledge discovery with machine learning tools, different combinations of statistical models and choices for predictor variables (feature selection) are tried to find combinations that can accurately predict the outcome variable(s) of interest. Those combinations that achieve high accuracy are scrutinized to discover the underlying patterns, some of which may lead to unexpected insights. While this search for accurate models is generally not incorporated into individual machine learning methods, there are some exceptions – for example, rule-discovery methods are designed to incorporate this search as part of the machine learning process.

Different machine learning methods are designed to identify different types of patterns. This is a ‘bias’ inherent in the type of model – only patterns anticipated by the form of the model can be discovered from the data. For example, models in the form of “if-then-rules” describe patterns in rectangular sub-spaces within the feature-space, thus this model is not optimal for detection of curved boundaries. Curved boundaries in the data may require other methods such as Support Vector Machines (SVM). Thus, automated knowledge discovery methods do not replace the human researcher – these tools must be guided by the understanding and experience of human researchers.

### **2.3.1 Machine Learning Concepts**

#### **2.3.1.1 Features and Attributes**

For learning from data, entities of interest (e.g. human subjects) are represented by a vector of feature-values – these ‘features’ (attributes of subjects) are observables that are used for prediction by the statistical model. For classification problems, the predicted variable is the ‘class’ label of the object (e.g. ‘disease’ or ‘normal’). For regression problems, the predicted variable is numeric (e.g. number of days spent in the hospital).

The machine learning problem can be visualized as learning the joint probability distribution of all the variables of interest (the predictor variables  $[x_1, x_2, \dots, x_p]$  and the predicted

variable  $y$ ). The data consists of a set of examples (e.g. human subjects) where each example is the tuple  $[x_1, x_2, \dots, x_p, y]$ . The  $k^{\text{th}}$  tuple in the dataset can be represented as  $[\vec{X}_k, y_k]$  where the vector  $\vec{X}_k$  represents the values of the predictor variables  $[x_1, x_2, \dots, x_p]$  for the  $k^{\text{th}}$  subject. Each example in the dataset is a point in the high-dimensional space spanned by the variables  $[x_1, x_2, \dots, x_p, y]$ . Thus, the dataset can be viewed as a cloud of points in this high-dimensional space. The goal of the machine learning method is to learn the distribution of these data-points in this space. The representation used by a particular machine learning method to represent this distribution is called the machine learning ‘model’. The models can be mathematical relationships between variables, descriptions of boundaries in the high-dimensional space, or explicit representations of the joint probability distribution.

### **2.3.1.2 Model selection and validation**

The motivation behind machine learning is to find statistical patterns in old data that can be generalized to new data from the same population. Thus, if a model of the joint probability distribution of old data correctly reflects the true joint probability distribution in the real-world, this model can be used for prediction of unknown values of variables. For example, a machine learning model created from fMRI data from existing patients may show that reduced neural activation in an anatomical region of the brain is a reliable predictor of Substance Use Disorder (SUD). This information is of clinical value only if reduced activation in this region is predictive of SUD for new patients. If a machine learning model is accurate for the dataset which was used to create the model (called the training set), but is inaccurate for a new dataset of comparable subjects (called the testing set), the model is not generalizable – this problem is referred to as ‘over-fitting’ or ‘over-training’. The ‘goodness-of-fit’ of a model refers to how well the model explains the data – the over-fitting problem is caused by achieving goodness-of-fit at the expense of generalizability.

Model selection refers to the task of choosing the model that is likely to be generalizable to new datasets. The generalizability of proposed models can be directly assessed by randomly subdividing the full dataset into two subsets – one subset is used for training the model and the other subset is used for testing (cross-validating) the model. Models are selected based upon their

cross-validation accuracy rates. Variations of the cross-validation approach include leave-one-out cross-validation and k-fold cross-validation.

Leave-one-out cross-validation involves repeated application of model-building with  $N-1$  examples ( $N$  is the total number of examples in the dataset) and model-testing against the example left-out during model creation. The average accuracy from  $N$  such models is interpreted as the expected accuracy for new data (if a similar model is created with all  $N$  examples). In  $k$ -fold cross-validation, the dataset is divided in  $k$  randomly selected subsets of equal sizes – the model is built with  $k-1$  subsets and tested with the remaining subset, repeating the process such that all subsets participate in model-building. Again, the average classification accuracy is interpreted as the expected accuracy for a new dataset, if a model is built from all the examples in the current dataset.

It is also possible to use model selection criteria based upon theoretical considerations. There are two aspects to model selection – the form of the model (e.g. linear regression vs. non-linear regression) needs to be chosen, and the parameters of the model (e.g. coefficients of regression) need to be estimated. The latter is typically estimated by choosing the set of parameters that maximizes the likelihood of observing the data – this is known as the Maximum Likelihood (ML) estimate. Likelihood is the probability of observing the actual dataset given a mathematical model of the data. Given model type  $M$ , parameter values  $\theta$  for the model, and assuming independence between the examples in the dataset, the likelihood  $\hat{L}$  of observing the collection of data tuples  $[y_k, \vec{X}_k]$  is computed as

$$\hat{L} = \prod_{k=1}^N p(y_k | \vec{X}_k, \theta, M) \quad (2.3.1)$$

For some models (e.g. linear regression), the ML estimate can be calculated in closed-form. For other models (e.g. non-linear regression), iterative search procedures are employed to estimate the ML parameters.

When choosing between models of different forms (with different model-complexity), choosing the model with highest overall likelihood can lead to over-fitting. To avoid this, when choosing between models, the likelihood is penalized by the complexity of the model. A typical expression used for this purpose is the Bayesian Information Criterion (BIC)

$$BIC = 2 \log(\hat{L}) - \log(N) k \quad (2.3.2)$$

where  $k$  is the number of free parameters in the model  $M$ , and  $N$  is the number of examples in the training set.

### **2.3.1.3 Feature construction and selection**

The accuracy of machine learning models is very dependent on the quality of the features (or attributes) incorporated into the model. Presence of irrelevant features in the model can degrade the performance of models – thus, feature construction and feature selection are critical aspects of building machine learning models. Feature construction refers to the creation of new variables based upon existing raw variables – for example, segmentation of fMRI activation images is a form of feature construction where attributes of segments are used for machine learning, instead of using attributes of underlying voxels in the image. Feature selection refers to the incorporation of only a subset of all available features into the machine learning model – a variety of subsets may be tried to identify the subset that yields the most accurate model. Typically, the features are sorted by some ‘interestingness’ measure and subsets of this sorted list are incorporated into the model.

### **2.3.2 Classical Statistical analysis vs. Machine Learning**

Classical statistics uses a framework of hypothesis rejection for scientific knowledge discovery from data – the goal of data analysis is to accept or reject a prior ‘null’ hypothesis that the data does not have the effect (relationship between variables) of interest. In other words, the null hypothesis states that the statistical model of the effect of interest is not valid. The ‘alternate’ hypothesis is that the effect of interest is present (or, the model of the effect is valid). For hypothesis testing, a test-statistic is computed from the observed data and the null-hypothesis is rejected or accepted based on the value of the test-statistic. Assuming that the null-hypothesis is true, the p-value is the probability of observing a value for the test-statistic, equally or more surprising (in the direction of the alternative hypothesis) than the value of the test-statistic actually observed. If this probability is too low (below an agreed upon critical threshold that controls the risk of mistakenly accepting the presence of an effect), the null hypothesis is rejected and the presence of the effect is accepted. It should be noted that the validity of the statistical inferences are dependent upon the validity of assumptions in the model, such as the absence of

un-modeled effects in the data and that the residual error terms are normally distributed. Also, if enough hypothesis tests are performed, some of the observed test-statistics will exceed the threshold corresponding to the critical p-value from chance alone (in the absence of any real effect). To guard against this risk of false discovery, the critical p-value is adjusted downward to account for multiple comparisons (e.g. by Bonferroni correction).

While this hypothesis-driven approach to knowledge discovery is effective for discovering relationships between variables, such knowledge is more useful for human interpretation than for utilization in computerized clinical tools. Thus, classical statistical analysis may suggest very strong relationship between two variables of interest (low p-value) but this by itself does not provide any direct mechanism for application of this knowledge in automated clinical tools. For such applications, it is more useful to have predictive knowledge models in a functional form that relates some variable  $y$  in terms of other variables  $\vec{X}$ ,  $y = f(\vec{X})$ . In classical statistics this is referred to as regression.

Machine learning methods also employ models in a functional form, including less conventional forms such as ‘if-then-rules’. In addition to predicting the value of the target variable, some learning techniques also provide a probability distribution  $p(y|\vec{X})$ , reflecting the degree of confidence in the prediction. The principal difference between classical regression and machine learning is in the form of the modeling function  $f(\vec{X})$ . While classical statistics restricts itself to functional forms that are mathematically tractable, machine learning models are less conservative – they include other forms of models such as ‘rules’ and ‘neural networks’ which cannot be subjected to closed-form analysis.

The specific machine learning methods used in this work are described in more detail in the following sub-sections.

### 2.3.3 Gaussian Naïve Bayes Classifier

The Naïve Bayes classifier [35] is based upon a generative model of the data – the learnt model can be used to generate an approximation to the dataset used to learn the model. For classification problems of the form  $y = f(\vec{X})$ , the model learns an approximation to the joint probability distribution  $p(y, \vec{X})$  such that given a set of values  $X_h$  for the predictor variables



$\vec{X}$ , it is possible to compute the conditional probability  $p(y = y_i | \vec{X} = X_h)$  for each class  $y_i$  of the target variable  $y$ . This is the posterior probability distribution for the class variable  $y$ , given the observed data.

From the definition of conditional probabilities, the joint distribution of the variables of interest is given by

$$p(y = y_i, \vec{X} = X_h) = p(\vec{X} = X_h | y = y_i)p(y = y_i) \quad (2.3.3)$$

where,  $p(\vec{X} = X_h | y = y_i)$  is the class conditional probability specifying the probability of observing the specific values  $X_h$  for the variables  $\vec{X}$ , given that class variable  $y$  has value  $y_i$ , and  $p(y = y_i)$  is the prior probability of observing the class  $y_i$ .

Using Bayes Rule, it is possible to compute the probability of interest  $p(y = y_i | \vec{X} = X_h)$ , which is the posterior probability of the class  $y_i$ , given the observed data  $X_h$ .

$$p(y = y_i | \vec{X} = X_h) = \frac{p(y = y_i, \vec{X} = X_h)}{p(\vec{X} = X_h)} = \frac{p(\vec{X} = X_h | y = y_i)p(y = y_i)}{p(\vec{X} = X_h)} \quad (2.3.4)$$

The marginal probability of the observed data  $p(\vec{X} = X_h)$  is computed by summing over all the possible choices ( $y_i$ ) for the class variable  $y$

$$p(\vec{X} = X_h) = \sum_i p(y = y_i, \vec{X} = X_h) \quad (2.3.5)$$

where,  $p(y = y_i, \vec{X} = X_h)$  is computed from the class conditional probability and the prior probability (using equation 2.3.3).

Thus Bayes classifiers allow the computation of the posterior probability of the class labels in terms of the class conditional probabilities and the prior probabilities of the individual class labels – these two latter probabilities must be estimated from the data.

The Naïve Bayes classifier makes a simplifying assumption about the class conditional probability  $p(\vec{X} = X_h | y = y_i)$ . It is assumed that the variables in  $\vec{X}$  are conditionally independent of each other given the class  $y_i$ . This permits computation of the joint class conditional probability  $p(\vec{X} = X_h | y = y_i)$  in terms of the individual class conditional probabilities:

$$p(\vec{X} = X_h | y = y_i) = \prod_j p(x_j = x_{j_h} | y = y_i) \quad (2.3.6)$$

where  $x_{j_h}$  is the observed value of for  $x_j$ , the  $j^{\text{th}}$  variable in  $\vec{X}$ .

In Gaussian Naïve Bayes, which is applicable when all the variables in  $\vec{X}$  are continuous variables, the individual class conditional densities are modeled with Gaussian distributions:

$$p(x_j | y = y_i) \sim N(\mu_{ij}, \sigma_{ij}) \quad (2.3.7)$$

Thus, for Gaussian Naïve Bayes classifiers, it is required to estimate the parameters  $\mu_{ij}$  and  $\sigma_{ij}$  of individual Gaussian class conditional densities  $p(x_j | y = y_i)$  from the training data. The prior probabilities  $p(y = y_i)$  are also estimated from the frequency of occurrence of individual classes in the training dataset. Once these are estimated, equation 2.3.4 can be used to compute the posterior probabilities for all possible classes, given any set of values for the variables  $\vec{X}$ . For classification purposes, the class with highest posterior probability is chosen as the predicted class.

### 2.3.4 Artificial Neural Networks

Artificial Neural Networks (ANNs) are networks of smaller processing units inspired by the architecture of biological networks of neurons. In the 1940s Hebb [36] suggested reinforcement of connections between neurons as a possible mechanism for neural learning. The Perceptron [37, 38] was the simplest instance of a feed-forward neural network used a linear classifier. Since then, development of multi-layer networks and the back-propagation algorithm [39] has popularized the application of ANNs in a wide variety of domains.

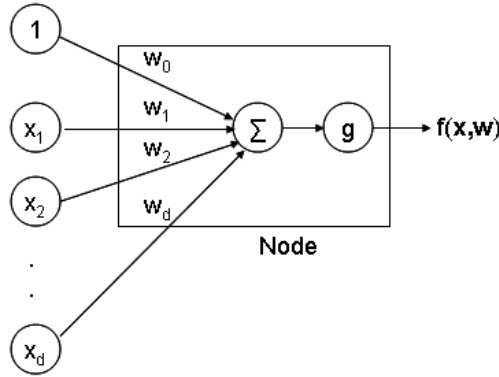
An ANN consists of an ensemble of ‘nodes’ and a set of connections between the nodes. Each node has a set of inputs and one output – the output-value of the node is based upon the input-values and a set of ‘weights’ maintained at the node. In a feed-forward ANN, the nodes are arranged in layers such that the nodes in the first layer accept the original input variables  $\vec{X}$ , and the output variable  $y$  is represented by the output of the single node in the last layer. The intermediate layers receive inputs ( $\mathbf{X}$ ) from nodes in the previous layer and propagate their outputs to the next layer.

The nodes can implement any function that maps the inputs (and a ‘bias’ term) to a real-valued output. For classification problems, a common choice is to use the logistic regression function at the node:

$$f(\mathbf{X}, \mathbf{W}) = g(w_0 + \sum_{j=1}^d w_j x_j) \quad (2.3.8)$$

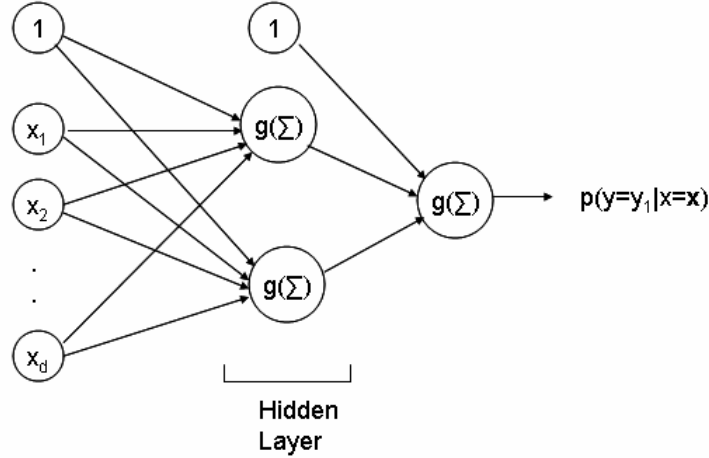
where  $\mathbf{W}$  is the set of weights ( $w_j$ ) for each of the  $d$  inputs ( $w_0$  is the bias term) and  $g(z)$  is the logistic (or sigmoid) function

$$g(z) = \frac{1}{(1 + e^{-z})} \quad (2.3.9)$$



**Figure 9.** A node in an Artificial Neural network.

An ANN (Figure 9) consists of an input layer with one input-node for each of the input variables (and a bias term), a set of ‘hidden’ or intermediate layers, and an output layer for the target variable. Each node (other than the input-nodes) has an input from all the nodes in the previous layer (Figure 10). For each node, the weights for each of the inputs are learnt from the training set. However, the architecture of network (number of nodes and layers) is typically chosen by the researcher based upon the complexity of the dataset.



**Figure 10.** Artificial Neural Network with one hidden layer with two nodes.

In an ANN, there is a weight associated with each of the arcs in the network. These weights are learnt from the training data – this is referred to as training the network. For a two-class classification problem, the goal of the training phase is to minimize the difference (error) between the actual class ( $y_i=0/1$ ) and the output of the network, interpreted as  $p(y = y_1 | \vec{X} = X_h)$ , given any set of values  $X_h$  for the input variables  $\vec{X}$ . For this purpose, the training phase searches for the set of weights that minimizes the prediction error for the training data. An iterative gradient search method [40] can be used for this purpose. Typically ‘online’ gradient search is used, where the weights are updated based on the prediction error as each of the training-set instances is encountered. If  $w_{ij}$  is the weight for the arc between nodes  $i$  and  $j$  (node  $i$  is at the level prior to that of node  $j$ ), the weight  $w_{ij}$  is updated based upon  $E_j$ , the prediction error-value at node  $j$ , and  $n_i$ , the output-value of the node  $i$

$$w_{ij} \leftarrow w_{ij} - \alpha E_j n_i \quad (2.3.10)$$

where  $\alpha$  is a learning rate that decays over training iterations to ensure convergence of the weights.

The prediction error-value for the last (output) node is calculated first, based upon the divergence between the true value ( $y_r$ ) of the target function and the value ( $\hat{y}_r$ ) predicted by the network for the  $r^{\text{th}}$  instance of the training set

$$E_o = -(y_r - \hat{y}_r) \quad (2.3.11)$$

The error-values for the hidden nodes are computed by propagating this error-value back through the network, one layer at a time. The error term for hidden node  $i$  is calculated from a weighted sum of error terms ( $E_l$ ) from nodes  $l$  in the next layer

$$E_i = [\sum_l E_l w_{il}] n_i (1 - n_i) \quad (2.3.12)$$

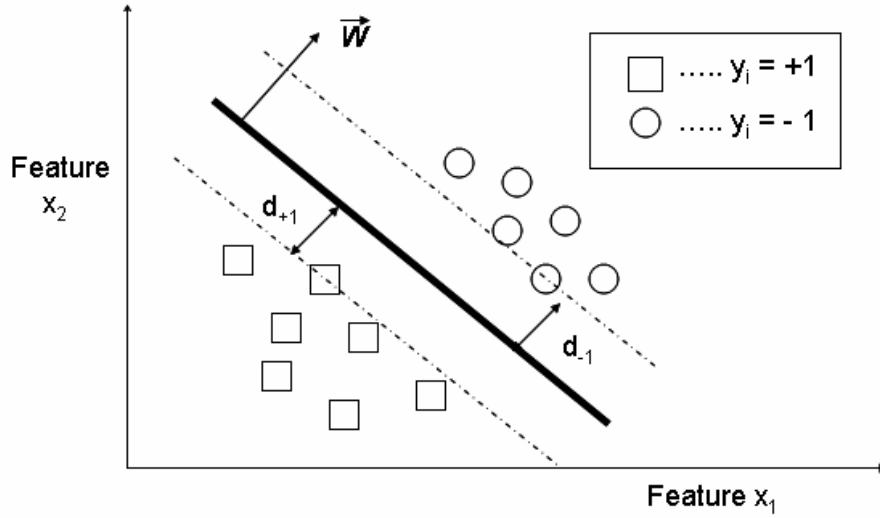
The weights of the network are typically initialized with random values before training. For training the network, the instances of the training set are presented one by one to the network (cycling to the beginning, at the end of the training set) and the weights are updated according to the above equations. This procedure is repeated for a pre-specified number of epochs (sweeps through the whole training set) or until the weights stop changing appreciably.

### 2.3.5 Support Vector Machine

Support Vector Machines (SVMs) [35, 41, 42] are classification methods that identify an optimal linear boundary (hyper-plane in feature-space) to discriminate between two groups. Though the separating boundary is linear, transformation of the original data vectors with ‘kernel’ functions can yield separating boundaries that are effectively non-linear in the original feature-space. Since many linear boundaries can separate the two groups, the optimal boundary is defined as the hyper-plane that maximizes the sum of the shortest distances from the boundary to instances of each class (this sum is called the ‘margin’). Since the separating surface is based upon instances of the training data that are on or near the margin – these data points are referred to as the ‘support vectors’.

Figure 11 shows the margin (with dotted lines), for a separating boundary. Here, it is assumed that the class variable  $y_i \in \{-1, +1\}$  and each example in the training set consists of  $[y_i, \vec{X}]$ , where  $\vec{X}$  is a vector of real-valued features. For each class, the minimum of the distances from the separating boundary are denoted as  $d_{+1}$  and  $d_{-1}$  respectively. The objective is to find the separating surface that maximizes this margin

$$M = d_{+1} + d_{-1} \quad (2.3.13)$$



**Figure 11.** The margin (dotted line) for SVM separating boundary.

If the classes are assumed to be separable by a linear boundary, it is possible to find a weight-vector  $\vec{W}$  and a bias  $w_0$  such that the plus-plane is represented by

$$\{x : \vec{W} \bullet x + w_0 = +1\} \quad (2.3.14)$$

and the minus-plane is represented by

$$\{x : \vec{W} \bullet x + w_0 = -1\} \quad (2.3.15)$$

where the weight-vector  $\vec{W}$  is perpendicular to both the plus-plane and the minus-plane.

Also, since the classes are assumed to be separable by such a linear boundary

$$\vec{W} \bullet \vec{X}_k + w_0 \geq 1 \text{ for all training examples } k, \text{ such that } y_k = 1 \quad (2.3.16)$$

and

$$\vec{W} \bullet \vec{X}_k + w_0 \leq -1 \text{ for all training examples } k, \text{ such that } y_k = -1 \quad (2.3.17)$$

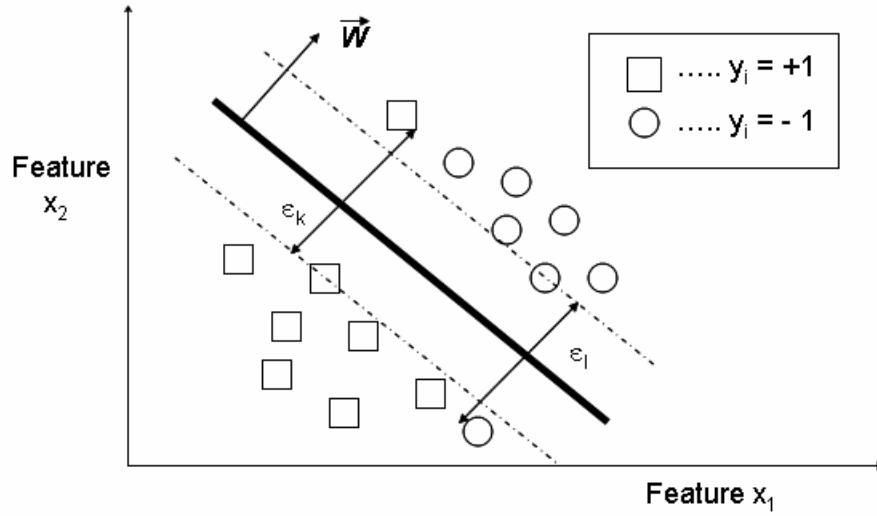
The width of the margin can be computed from the weight-vector

$$M = d_{+1} + d_{-1} = \frac{2}{|\vec{W}|} = \frac{2}{\sqrt{\vec{W} \bullet \vec{W}}} \quad (2.3.18)$$

The optimal separating boundary is given by the set of weights  $\vec{W}$  that maximizes the margin, which is equivalent to minimizing  $\vec{W} \bullet \vec{W}$  subject to  $N$  constraints, which require that each of the examples in the training set is correctly classified. This problem can be solved by

Quadratic Programming, a class of optimization algorithms that maximizes a quadratic function of some real-valued variables, subject to some linear constraints (equalities and inequalities).

Since the assumption of linear separability of the classes is un-realistic, error terms are introduced to penalize for misclassification. If a data point is misclassified by the separating boundary, the assessed penalty (shown in Figure 12) is based on the distance from the misclassified data point to the plane of correct classification (plus-plane or minus-plane).



**Figure 12.** Penalty terms ( $\hat{U}$ ) for misclassification by SVM.

With these penalties for misclassification, the objective function ( $J$ ) to minimize includes  $N$  error terms ( $\varepsilon_i$ ) in addition to the weights ( $\vec{W}$ ) and the bias term ( $w_0$ )

$$J(\vec{W}, w_0, \vec{\mathcal{E}}) = \frac{\vec{W} \bullet \vec{W}}{2} + C \sum_{i=1}^N \varepsilon_i \quad (2.3.19)$$

where, the user-specified parameter  $C$  controls the penalty for each misclassification.

There are  $2N$  constraints:

$$\vec{W} \bullet \vec{X}_k + w_0 \geq 1 - \varepsilon_k \text{ for all } k, \text{ such that } y_k = 1 \quad (2.3.20)$$

and

$$\vec{W} \bullet \vec{X}_k + w_0 \leq -1 + \varepsilon_k \text{ for all } k, \text{ such that } y_k = -1 \quad (2.3.21)$$

and

$$\varepsilon_k \geq 0 \text{ for all } k. \quad (2.3.22)$$

The same quadratic programming problem can be alternatively formulated with Lagrange multipliers ( $\alpha_k$ ), where the objective function is

$$J(\alpha_k) = \sum_{k=1}^N \alpha_k - \frac{1}{2} \sum_{k=1}^N \sum_{l=1}^N \alpha_k \alpha_l Q_{kl} \quad (2.3.23)$$

where  $Q_{kl} = y_k y_l (\vec{X}_k \bullet \vec{X}_l)$

subject to the  $2N$  constraints

$$0 \leq \alpha_k \leq C \text{ for all } k \quad (2.3.24)$$

and

$$\sum_{k=1}^N \alpha_k y_k = 0 \quad (2.3.25)$$

From the optimal solution  $\{\alpha_1, \alpha_2, \dots, \alpha_N\}$  to the above quadratic programming problem, the weight vector and the bias-term for the separating boundary can be computed as

$$\vec{W} = \sum_{k=1}^N \alpha_k y_k \vec{X}_k \quad (2.3.26)$$

and

$$w_0 = \frac{\max_{i \text{ s.t. } y_i = -1} (\vec{W} \bullet \vec{X}_i) + \min_{i \text{ s.t. } y_i = 1} (\vec{W} \bullet \vec{X}_i)}{2} \quad (2.3.27)$$

The final classification function is

$$f(\vec{X}, \vec{W}, w_0) = \text{sign}(\vec{W} \bullet \vec{X} + w_0) \quad (2.3.28)$$

While SVMs are designed to find the optimal *linear* boundary, ‘basis’ functions can be used for non-linear separation. Basis functions are used to redefine the feature space – extended-features  $\vec{\chi}$  are constructed from the original features  $\vec{X}$  such that linear boundaries in the extended-feature-space are non-linear in the original feature-space.

$$\vec{\chi} = \Phi(\vec{X}) \quad (2.3.29)$$

For example, if  $\vec{X} = [x_1, x_2]^T$ , the extended-features can be constructed as

$$\vec{\chi} = \Phi(\vec{X}) = [x_1^2, x_2^2, \sqrt{2} x_1 x_2, \sqrt{2} x_1, \sqrt{2} x_2, 1]^T \quad (2.3.30)$$



Since, in the formulation of the quadratic programming problem with Lagrange multipliers, the training data appears only as dot products between examples (see equations 2.3.23-2.3.28), the search for the optimal boundary can be performed in the extended feature-space by replacing  $(\vec{X}_k \bullet \vec{X}_l)$  with  $(\Phi(\vec{X}_k) \bullet \Phi(\vec{X}_l))$ . Linear discrimination in the extended feature-space corresponds to non-linear discrimination in the original feature-space.

A Kernel is a function that returns the dot-product between variables in the extended-feature-space

$$K(\vec{X}_k, \vec{X}_l) = (\Phi(\vec{X}_k) \bullet \Phi(\vec{X}_l)) \quad (2.3.31)$$

With proper selection of basis functions, the kernel function  $K(\vec{X}_k, \vec{X}_l)$  can be computed in terms of the original input variables  $\vec{X}_k$  and  $\vec{X}_l$ . For example, for the above quadratic basis-expansion (equation 2.3.30),

$$K(\vec{X}_k, \vec{X}_l) = (1 + (\vec{X}_k \bullet \vec{X}_l))^2 \quad (2.3.32)$$

Thus, for SVM learning in the extended-space, only the Kernel function needs to be specified (not the actual basis functions used for feature-expansion). The SVM implementation [43] used in this work employs the following “radial basis function” as the kernel

$$K(\vec{X}_k, \vec{X}_l) = e^{-\frac{1}{p} |\vec{X}_k - \vec{X}_l|^2} \quad (2.3.33)$$

where  $p$  is the dimension of the data vector  $\vec{X}$ .

## 2.4 MACHINE LEARNING FROM ACTIVATION MAPS

The vast majority of fMRI experiments are analyzed with methods based upon classical statistics (section 2.2) – the results of experiments are reported as co-ordinates (and p-values) of activation corresponding to particular tasks (and between-group differences, if any). However, for clinical applications, automated classification models are more useful than p-values. For this purpose, machine learning techniques have been employed to build classification models from sets of activation maps (SPMs). For machine learning, a subset of the voxels from the activation maps

can be used directly as features in the machine learning model, or features can be constructed from the activation images. Ford et al. [2] have used Principal Component Analysis (PCA) for feature-construction from fMRI activation maps to distinguish between patients and controls for Alzheimer’s disease, schizophrenia, and concussions. Using voxel-based feature construction from activation maps, Zhang et al. [44] have the studied the feasibility of distinguishing between healthy controls and patients with Substance Use Disorders (SUD). These approaches to feature construction from activation images are described next.

### 2.4.1 Feature construction and selection

Since fMRI activation images consist of larger number of voxels (typically over 100,000) and relatively smaller number of subjects (typically 10-30), it is not feasible to use all the voxels in the machine learning model – large number of features increases the risk of over-fitting. Two feature-reduction techniques have been proposed for activation-image datasets –Principal Component Analysis (PCA) [2, 45] and  $k$ -best voxels (KBV) [44].

#### 2.4.1.1 PCA

PCA [46] is a standard method for dimensionality reduction that creates a smaller number of uncorrelated variables from linear combinations of correlated variables. It is mathematically equivalent to a matrix decomposition known as singular value decomposition (SVD). The application of SVD to a set of activation images require the images to be ‘flattened’ to a  $P$ -dimensional vector, where  $P$  is the total number of voxels in each image (for example, this flattened layout can be the same as the linear storage layout of 3D arrays in computer memory). Thus, a collection of  $N$  activation images  $I_1, I_2, \dots, I_N$  is represented as the matrix  $M$  with  $N$  rows and  $P$  columns

$$M = \begin{bmatrix} I_1 \\ I_2 \\ . \\ . \\ I_N \end{bmatrix} \quad (2.4.1)$$

The singular value decomposition of  $M$  is

$$M = A D \Psi^T \quad (2.4.2)$$

where

$$A = \begin{bmatrix} a_1 & a_2 & \cdots & a_R \\ \downarrow & \downarrow & & \downarrow \end{bmatrix},$$

$$D = \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_R \end{bmatrix},$$

and

$$\Psi = \begin{bmatrix} \Psi_1 & \Psi_2 & \cdots & \Psi_R \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}$$

Here  $R$  is the rank of the matrix  $M$ ,  $a_i$  are the  $N$ -dimensional left-eigenvectors of  $M$ ,  $D$  is a  $R$ -by- $R$  diagonal matrix of the eigenvalues of  $M$ , and  $\Psi_i$  are the  $P$ -dimensional right-eigenvectors of  $M$  (the un-flattened  $\Psi_i$  vectors are also known as eigenimages, see Figure 36 for an example). Thus, the matrix  $A$  is a representation of the original images in terms of new basis vectors (eigenimages). For typical image datasets, the rank of the matrix  $M$  is

$$R = \min(N, P) = N \quad (2.4.3)$$

Thus, the  $N$ -by- $N$  matrix  $A$  represents a reduction of the original  $N$ -by- $P$  matrix  $M$ , which corresponds to a reduction of the  $P$ -dimensional voxel-based feature-space to an  $N$ -dimensional eigenimage-based feature-space. While the original  $P$ -dimensional feature-space considers the voxels as individual features, the eigenimages can be considered as a mechanism for grouping voxels for feature construction (somewhat similar to segmentation). However, since  $R$  is limited by the number of subjects  $N$ , this grouping may not be sensitive enough to isolate small pockets of voxels that may be differentially activated between groups of subjects (normal/disease). Small pockets of voxels that exhibit differential activation between groups of subjects may get incorporated into larger eigenimages (larger grouping of voxels) which may not as effective for discrimination.

For feature selection, the eigenimage features are added to the machine learning model one at a time, ordered by the magnitudes of the eigenvalues (representing the importance of the corresponding eigenimage in explaining the variance in the data). Thus, inclusion of fewer

features corresponds to a coarse representation of the image data, while inclusion of all the features may incorporate noise elements that are not relevant for discrimination.

#### 2.4.1.2 *k*-best voxels (KBV)

In the *k*-best voxels (KBV) approach, the voxels are ranked by some ‘interestingness’ criterion and a certain number of the ‘best’ voxels are retained as features in the classification model. For example, voxels that exhibit highest inter-group differences in activation levels may be chosen as features. Thus, if the activation level for the  $i^{\text{th}}$  voxel from the  $n^{\text{th}}$  image in the ‘normal’ group is denoted by  $I_{ni}$  and the activation level for the same voxel from the  $d^{\text{th}}$  image in the ‘disease’ group is denoted by  $I_{di}$ , an interestingness measure for the  $i^{\text{th}}$  voxel can be computed as

$$T_i = \text{abs}(ttest(\{I_{ni} \mid n \in \{k \mid y_k = 0\}\}, \{I_{di} \mid d \in \{k \mid y_k = 1\}\})) \quad (2.4.4)$$

where  $\{k \mid y_k = 0\}$  denotes the indices for the ‘normal’ images,  $\{k \mid y_k = 1\}$  denotes the indices for the ‘disease’ images and  $ttest()$  denotes the two-sample t-test (equation 2.2.1). Since this measure can be vulnerable to false-positives arising from noise levels outside the boundary of the brain in the activation images, an alternative measure based upon the absolute difference between group-means was adopted as the interestingness measure for this work.

$$D_i = \text{abs}\left(\frac{\sum_{n \in \{k \mid y_k = 0\}} I_{ni}}{|\{k \mid y_k = 0\}|} - \frac{\sum_{d \in \{k \mid y_k = 1\}} I_{di}}{|\{k \mid y_k = 1\}|}\right) \quad (2.4.5)$$

where  $D_i$  is the interestingness measure for the  $i^{\text{th}}$  voxel.

Note that spatial relationships between the selected voxels are not considered in the KBV approach to feature selection – selected voxels may be spatially grouped or isolated from each other. Chapter 6 presents a refinement to the KBV approach that incorporates spatial coherence information in the feature selection stage to improve classification accuracy.

### 2.4.2 Motivations for new approaches

As discussed earlier, with current tools, precise spatial normalization of brains of different subjects is a difficult task. The key assumption inherent in current feature selection methods

(PCA, KBV) is that voxels are comparable across subjects – that is, knowledge discovery is voxel-centric. Smoothing of images to account for imprecise spatial normalization is of questionable efficacy, and it leads to loss of spatial resolution. Also, the conventional approaches to feature construction (PCA, KBV) do not provide a direct mechanism for discovering knowledge about sizes of activated regions.

To address these concerns, the KDSf framework proposes construction of features based upon characteristics of functional segments – thus, knowledge discovery is segment-centric. The key assumption is that clumps of voxels that exhibit hemodynamic co-modulation belong to a spatially localized functional unit of the brain. Given this assumption, it is hypothesized that machine learning from the characteristics of these functional units can be more effective than conventional methods of feature construction.

Very recently, Thirion et al. [47] have also proposed a segment-centric method for analysis of single-group fMRI datasets. In their method, for identification of consensus regions of activation in a group of subjects, parcel-based random effects analysis was shown to be more sensitive than standard voxel-based random effects analysis. The parcel-based random effects analysis uses the average of the estimated effects from all the voxels that belong to the same ‘clique’, which is the set of equivalent (but possibly misaligned) functional segments across subjects. While their work currently does not address classification between groups, such an extension is feasible. The main difference between their approach and the KDSf approach is that KDSf does not restrict itself to univariate statistics based upon average activation strengths for knowledge discovery. In KDSf, the image dataset is reduced to a standard machine learning table where all the characteristics of the functional segments (size, shape, location, activation strength etc.) are considered for knowledge discovery. This facilitates the search for different classification models (with possibly multivariate feature subsets) that can adequately discriminate between the groups.

The KDSf framework requires a mechanism for automatic isolation of such functional units – the ACEIC [9] method is proposed for this ‘functional segmentation’. Also, functional segments from different subjects need to be ‘registered’ to permit comparison of their characteristics across subjects. For this purpose, the KDSf framework currently incorporates a greedy segment-registration strategy (see Appendix). While this simple greedy registration strategy does not use anatomical information, more advanced segment-registration schemes can

also be accommodated by the KDSf framework. Finally, with features constructed from registered functional-segments, the KDSf framework uses off-the-shelf machine learning methods for knowledge discovery.

## **2.5 CONVENTIONAL FUNCTIONAL SEGMENTATION**

Three main approaches have been proposed for grouping voxels based on similarity of timecourses – clustering [48-51], Independent Component Analysis [52-54] and image segmentation methods [55]. Of these, the first two approaches do not restrict the voxels within a cluster (or component) to be spatially connected.

### **2.5.1 Clustering**

A wide variety of clustering methods, including k-means [49], hierarchical clustering [48], and fuzzy clustering [50] have been employed for grouping voxels based upon similarity of their timecourses. In general, the goal of cluster analysis [56, 57] is to find groups such that objects in a group are as similar as possible and the objects in different groups are as dissimilar as possible. The objective of cluster analysis is to *discover* groups (or structure in the data), whereas the objective of classification methods is to assign objects to groups that were defined in advance (in the training set). Note that the discovered clusters are meaningful only in the context of the features employed in the analysis – the use of a different set of features will likely produce a different grouping for the same set of objects. Also, clustering results are dependent on the distance function used to evaluate dissimilarity between objects represented by a set of features.

#### **2.5.1.1 Distance Metrics for Clustering**

Clustering requires the specification of a distance metric to evaluate the similarity (and dissimilarity) between objects. For numerical data, some of the choices for distance metric are listed below.

Euclidean distance

This distance metric is the geometric distance between points representing the objects in feature-space. The feature-space is a  $p$ -dimensional space where each feature is represented by an axis. The feature vector for the  $i^{\text{th}}$  object is represented as  $X_i = (x_{i1}, x_{i2}, \dots, x_{ip})$  and the Euclidean distance between objects  $i$  and  $j$  is given by

$$d(i, j) = ((X_i - X_j) \bullet (X_i - X_j)^T)^{1/2} \quad (2.5.1)$$

#### City-block or Manhattan distance

This distance metric is sum of the distances along individual axes in the feature space. This distance can be used where the total sum of individual differences is meaningful.

$$d(i, j) = \sum_{k=1}^p |x_{ik} - x_{jk}| \quad (2.5.2)$$

#### Minkowski distance

This is also known as the  $L_q$  metric. Euclidean distance ( $q=2$ ) and Manhattan distance ( $q=1$ ) are special cases of this metric.

$$d(i, j) = (|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \dots + |x_{ip} - x_{jp}|^q)^{\frac{1}{q}} \quad (2.5.3)$$

where  $q$  is any real number larger than or equal to 1.

#### Correlation distance

This distance metric can be effective for evaluating the similarity of ‘shapes’ of time-courses even when the time-courses have different amplitudes.

$$d(i, j) = 1 - \frac{(X_i - \overline{X_i})(X_j - \overline{X_j})^T}{[(X_i - \overline{X_i})(X_i - \overline{X_i})^T]^{1/2} [(X_j - \overline{X_j})(X_j - \overline{X_j})^T]^{1/2}} \quad (2.5.4)$$

where  $\overline{X_i} = \frac{1}{p} \sum_k x_{ik}$

### **2.5.1.2 Clustering Methods**

#### k-means clustering

This is a partitioning method where the number of partitions ( $k$ ) is specified in advance by the user. The centroid of the cluster is defined as a point in  $p$ -dimensional space found by averaging the coordinates of the objects in the cluster along each dimension. However, the location of the centroid of a cluster need not coincide with the location of any of the objects. The

objective is to minimize the sum of squared distances of cluster members from the corresponding centroids. Thus this method can be considered to be a variance minimization technique.

A simple version of the  $k$ -means algorithm can be stated as follows [56]:

1. Start with a random partition (with  $k$  clusters) and compute the  $k$  centroids for the clusters.
2. Assign all objects to the cluster with the strictly nearest centroid. If no object was re-assigned, then stop.
3. Re-compute the centroids for clusters. Go to step 2.

Since the objects move between clusters only when the distance from the centroid is reduced, the sum of squared distances must decrease with each move. This guarantees convergence to a locally optimal solution. However, since the cluster is represented by a centroid location (and not an object), some clusters can lose all their members, leading to fewer than  $k$  clusters.

The  $k$ -means method is sensitive to the initial choice of cluster-partition and multiple runs may be needed to verify the robustness of the clustering solution. The solution may also depend on the order in which objects are presented to the algorithm.

### Hierarchical Clustering

The hierarchical methods do not construct a partition with a pre-specified number of clusters. Instead, it provides a hierarchical representation of all possible choices for  $k$ . The user can choose  $k$  based upon the hierarchical representation – for example, distance-based criteria can be used to separate the branches of the hierarchy at some level of the tree. There are two classes of hierarchical techniques – agglomerative and divisive. In the agglomerative technique, individual objects start as singleton clusters and coalesce with other clusters to yield bigger clusters, eventually producing a cluster containing all the objects. At each step, some criterion is used to choose the next two clusters to merge. In a divisive method, the process runs in the reverse direction – from bigger clusters to smaller clusters. The results of agglomerative clustering can be different from divisive clustering.

The greedy nature of the hierarchical methods (agglomeration or division decisions are never undone), can lead to sub-optimal solutions. Since the partitioning approaches (e.g.  $k$ -means) try to find the optimal solution with  $k$  clusters, solutions found by partitioning



approaches can outperform solutions found by hierarchical approaches. However, the hierarchical methods allow the user to specify distance-based criteria for cluster purity – for example, the heterogeneity of fMRI timecourses within a segment can be controlled during functional segmentation.

The process of Agglomerative Hierarchical Clustering [58] is as follows:

1. Start by assigning each item to its own cluster. The distances between the clusters are initialized to the distances between the items in the clusters.
2. Find the closest pair of clusters and merge them into a single cluster.
3. Compute distances between the new cluster and each of the remaining clusters from step 2.
4. If all items are in a single cluster, stop. Else, go to step 2.

In this method, all objects start as their own singleton clusters. In each successive step, the closest clusters are merged to yield another cluster. The determination of the closest clusters requires computation of distances between clusters. There are several possibilities for evaluation of inter-cluster distances:

*Un-weighted Pair Group Method with Arithmetic mean (UPGMA)*

In this case, the distance between clusters  $A$  and  $B$  is taken as the average of the distances  $d(i,j)$  where object  $i$  is from  $A$  and object  $j$  is from  $B$ . As the clusters are merged, the distance between the merging clusters is a monotone function. That is, the critical (smallest) distance at any stage is higher than the critical distances from the earlier stages (prior merges).

For the UPGMA method, the distance measures need to be updated when two clusters  $A$  and  $B$  are merged to form a new cluster  $R$ . However, the average distance of  $R$  to other clusters  $Q$  can be computed without considering all the members of the clusters  $R$  and  $Q$ .

$$d(R, Q) = \frac{|A|}{|R|} d(A, Q) + \frac{|B|}{|R|} d(B, Q) \quad (2.5.5)$$

The use of this update relationship considerably improves the computational and memory requirements of the algorithm.

*Nearest Neighbor (or Single Linkage)*

For nearest-neighbor linkage, the between-cluster distance  $d(R, Q)$  is defined as the smallest of the pair-wise distances between the members of clusters  $R$  and  $Q$ . This can cause many objects to chain together into clusters that are not very homogeneous. However, this method is useful for detecting elongated clusters which cannot be detected by other methods. The single linkage method is equivalent to using the Minimum Spanning Tree of the graph of objects.

*Farthest Neighbor (or Complete Linkage)*

In complete linkage,  $d(R, Q)$  is defined as the maximum of the pair-wise distances between the members of clusters  $R$  and  $Q$ . This method tends to produce compact clusters. However the clusters may not be well separated. The ACEIC method uses this approach to enforce cluster homogeneity – clusters are merged as long as the maximum pair-wise distances between cluster members (timecourses) fall within a threshold.

Other linkage schemes include Centroid Linkage, where the clusters with the minimum Euclidean distance between the two centroids are merged. In Ward’s method the goal is to choose the cluster-pair for merger for which the increase in the error sum of squares (ESS) is the minimum.

One problem with clustering methods is the need for user-specified parameters (e.g. initial number of clusters or level of the cut in the hierarchy) which determine the quality of the clustering results. Recently, a comparative analysis of various clustering methods for timecourse data has shown that the accuracy of clustering results is typically dependent upon the appropriate choice of parameters and upon random initializations [51]. The ACEIC method (Chapter 5) uses iterative clustering with contrast maximization to avoid user-specified parameters.

## **2.5.2 Independent Component Analysis (ICA)**

Independent Component Analysis (ICA) is a method for recovering original signals from linear mixtures of these signals. Higher order statistics are used to find a set of signal ‘components’ that are maximally independent of each other [59, 60]. For a set of fMRI timecourses, it is possible to find spatially or temporally independent components – however, since the spatial dimension is much larger (hundred of thousands of voxels) than the temporal dimension (hundreds of time-

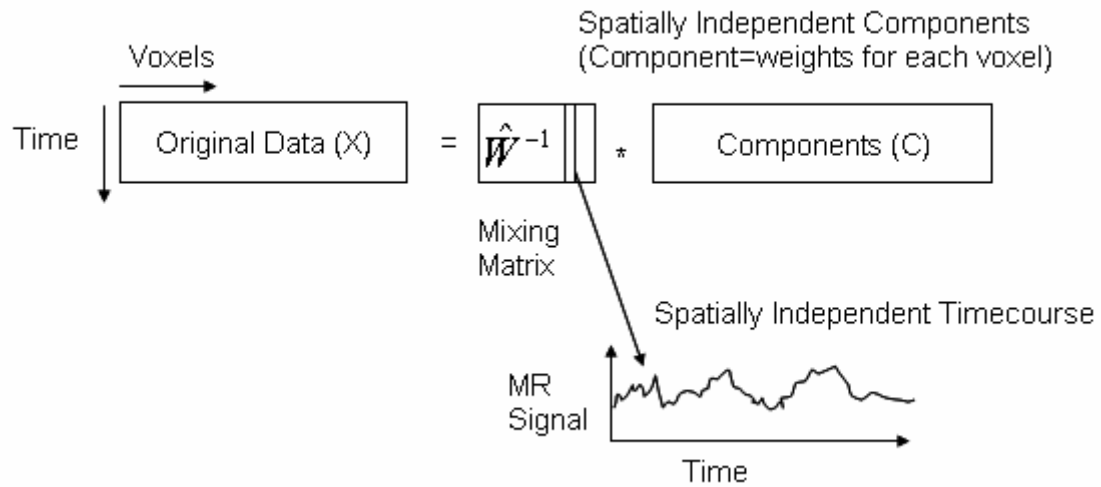
points), separation into spatially independent components is less demanding in terms of the sizes of the matrices to be manipulated [61].

For spatial ICA of fMRI, the data matrix ( $X$ ) has  $N$  rows (one for each time-point) and  $M$  columns (one for each voxel). The ‘signals’ to be separated are ‘flattened’ instances of image volumes, represented as  $M$ -dimensional vectors (Figure 13). The ICA decomposition of  $X$  into spatially independent components  $C$  is given by

$$C = \hat{W} X \quad (2.5.6)$$

where  $\hat{W}$  is the  $N$ -by- $N$  un-mixing matrix estimated by ICA and  $C$  is the  $N$ -by- $M$  matrix containing the  $N$  independent components. The estimated mixing matrix is  $\hat{W}^{-1}$ , and the original timecourse data can be reconstituted from the components by

$$X = \hat{W}^{-1} C \quad (2.5.7)$$



**Figure 13.** Schematic of spatial ICA of fMRI data

While Principal Component Analysis (PCA) identifies components that are un-correlated with each other, the ICA methods impose a stronger ‘independence’ criterion upon the components [61]. Independence requires that the generalized co-variance between two variables  $x$  and  $y$  is zero for all integers  $p$  and  $q$ :

$$C_{pq}(x, y) = E_{x,y}[x^p y^q] - E_x[x^p] E_y[y^q] = 0 \quad (2.5.8)$$

where  $E_x[\cdot]$  is the expectation over the probability distribution of  $x$ ,  $E_y[\cdot]$  is the expectation over the probability distribution of  $y$ , and  $E_{x,y}[\cdot]$  is the expectation over the joint probability distribution of  $x$  and  $y$ .

PCA only enforces that the variables are un-correlated:

$$C_{11}(x, y) = 0 \quad (2.5.9)$$

Each independent component is a vector of weights (one weight for each voxel) that specifies the degree of participation of the voxel in the component – thus, this is a mechanism for grouping voxels based upon a shared sub-pattern in the timecourses for the voxels. The voxels that exhibit ‘significant’ participation in a component (i.e. exhibit significant presence of a temporal sub-pattern) are identified by thresholding the component weight-vector with a user-specified threshold. An example of results from ICA of fMRI data is shown in Figure 49. There are several algorithms that can be used for estimation of independent components including kurtosis based methods [59] and information-theoretic methods [60].

Probabilistic ICA [53] allows for non-square mixing with Bayesian estimation of the model order (number of spatial components). In addition, in Probabilistic ICA (PICA), the independent component maps are assessed for significance with a Gaussian mixture model, without requiring the user to specify a threshold. However, the accuracy of ICA (and PICA) is dependent upon the validity of the model assumptions [61].

In this work, the accuracy of the ACEIC method for functional segmentation is compared with that of MELODIC [62], a publicly available implementation of Probabilistic ICA.

### 2.5.3 Image processing methods

Image processing approaches such as seeded region-growing have been proposed for functional segmentation of fMRI data [55]. In seeded region-growing, starting with a seed voxel, neighbor voxels are added to the region (segment) definition until some stopping criterion is reached – typically a region homogeneity criterion (RHC) is used as the stopping criterion. The region homogeneity criterion requires that a homogeneity measure computed for the time-courses in the region (segment) satisfies a user-defined threshold. However, given the variety of noise sources in fMRI data, it is difficult to choose the optimal homogeneity threshold for individual images. Other typical image segmentation approaches, such as split-and-merge segmentation [63], are also dependent upon the proper choice of region homogeneity thresholds.

Greedy region-growing with maximization of region-contrast has been used for intensity-valued medical images [64]. In this approach, starting with a seed pixel in the segment, the

neighboring pixel that is most similar to the pixels already in the segment is added to the segment. This process is repeated to optimize region-contrast as computed from intensities of voxels inside and outside the segment. While contrast-maximization does not require user-specified parameters, contrast measures used with intensity-valued images may not be suitable for timecourse-valued images. Further, contrast-maximization with greedy region-growing can lead to under-segmentation for fMRI image regions with lower levels of neural activation (see Chapter 5).

The development of ACEIC is motivated by the desire for an accurate segmentation method that does not require dataset-specific parameters such as number of clusters or homogeneity thresholds. ACEIC employs maximization of region-contrast along with a generalization of greedy region-growing to achieve these goals.

### **3.0 ANNOTATED EXAMPLE**

The KDSf framework is designed for automated knowledge discovery about cortical activation patterns that can be useful for discrimination between groups of subjects. For example, during performance of a task, subjects with a particular disease may exhibit higher activation levels at a specific cortical location. The goal of the KDSf framework is to automatically discover such differences even though such differences may be difficult to detect by visual inspection or voxel-by-voxel analysis. This chapter presents an example of an fMRI experiment that is designed to detect such differences between two groups of subjects. The steps of the knowledge discovery process are illustrated with the help of this example.

#### **3.1 CLINICAL APPLICATION: SUBSTANCE USE DISORDER**

Functional neuro-imaging provides a mechanism for observing the neurological underpinnings of Substance Use Disorder (SUD). While it is possible to assess risk for SUD with psychological tests, empirical evidence about differences in cortical activations associated with SUD is also emerging [65-67]. For example, a study which examined fMRI response in adolescents with history of alcohol and marijuana use observed less activation in inferior frontal and temporal regions and more activation in dorsolateral prefrontal cortex (DLPFC) when compared to controls [67].

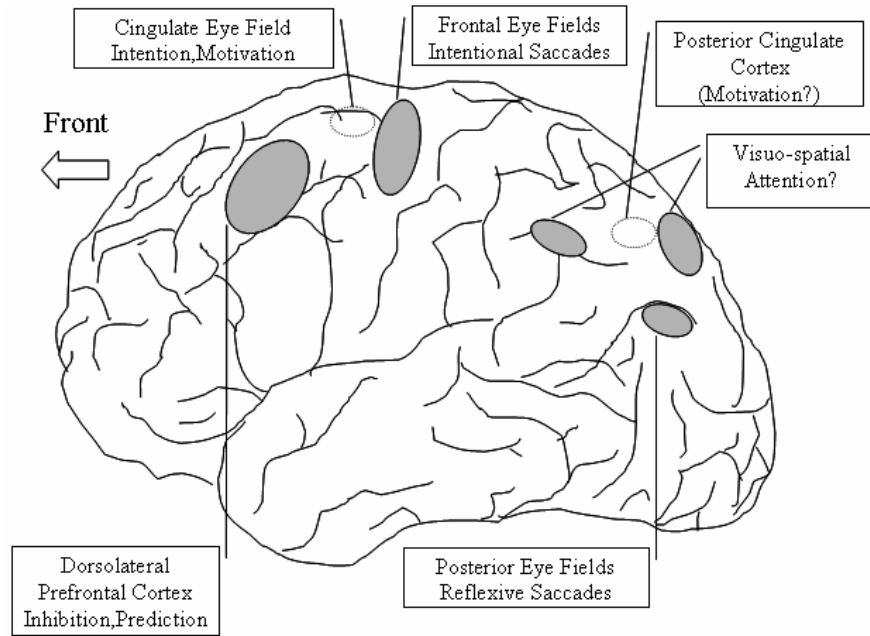
The risk for SUD is generally associated with impaired regulation of inhibitory processes. A trait termed ‘neurobehavior disinhibition’ (ND), which is derived using measures of executive function, affect modulation, and behavioral control, can effectively discriminate between youths at high and low risk for SUD [68]. The ND score can also significantly predict SUD between late childhood and young adulthood. While the ND is based on psychological

assessments, the neuro-anatomical correlates of this trait are not well understood. Functional imaging provides the opportunity to determine whether the ND-score can be associated with differences in cortical activation patterns. Functional imaging has the potential to identify the cortical correlates of inhibitory impairments and also to track the impact of therapeutic measures on these cortical patterns.

The goal of this fMRI study is to compare activation patterns between youths at high and low risk for SUD. The cortical activation patterns induced by an eye-tracking task are compared for two groupings of adolescents based upon their ND score – these two groups of subjects are referred to as ‘lowND’ and ‘highND’. The experimental paradigm and data analysis methods are described below.

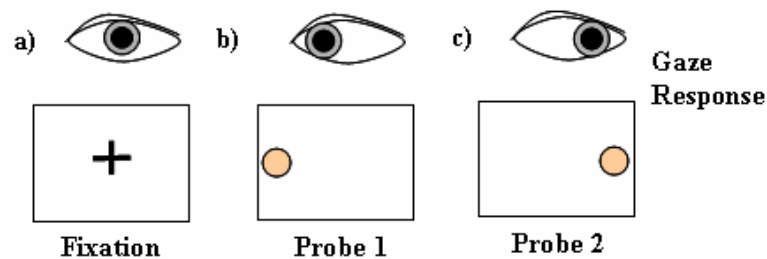
### **3.1.1 Experimental Paradigm**

Two eye-movement paradigms are employed to assess the ability of the subjects to inhibit a reflexive eye movement response to a visual stimulus. The first experimental paradigm compares visually guided saccades (VGS) against fixation, and the second experimental paradigm compares anti-saccadic eye movements against pro-saccadic eye movements. The VGS *task* is designed to establish integrity of saccadic control, whereas the anti-saccade *task* targets brain processes involved in suppression of the reflexive response (inhibition). The main cortical areas involved in saccadic control are shown in Figure 14 (adapted from [69]).



**Figure 14.** Main cortical areas (‘functional units’) involved in saccadic control (adapted from [69]).

The VGS task consists of 30 seconds of fixation on a white cross-hair in the middle of a screen alternated with 30 seconds of saccadic eye movements. During the saccadic eye movement block, dots of light appear horizontally to the left or right of the cross-hair location. The subjects are instructed to turn their gaze towards each dot as it appears (Figure 15). Thus the VGS *task* compares two *conditions* – pro-saccade vs. fixation. Each ‘run’ consists of six blocks of fixation alternated with six blocks of pro-saccade. The task timecourse can be represented as a ‘box-car’ function alternating between 30 second blocks of zeros (representing fixation) and 30 second blocks of ones (representing pro-saccade).

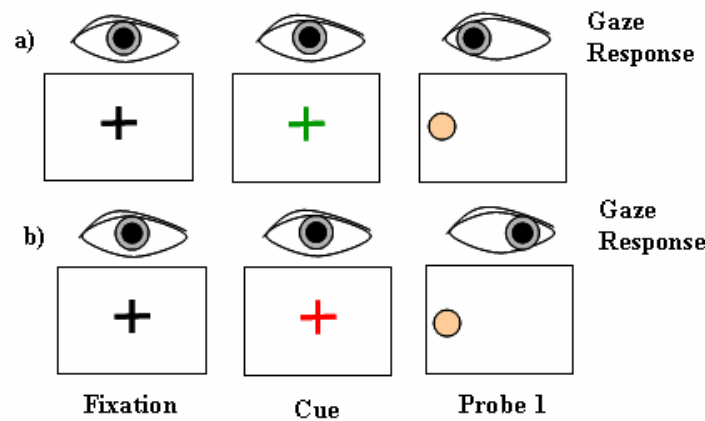


**Figure 15.** The VGS task alternates between 30 seconds blocks of fixation (a) and pro-saccade (b and c). During the pro-saccade task, the subject’s gaze moves towards the dot when the probe is presented. During the pro-saccade block, the probe randomly alternates between Probe 1 and Probe 2.

For the anti-saccade task (Figure 16), subjects are instructed to fixate on a white cross hair in the center of the screen. For each *trial*, the cross hair turns either green or red to indicate



the type of trial – pro-saccade or anti-saccade [70]. For a pro-saccade trial (cued by green cross-hair) the subjects are instructed to look at a dot that appears in the horizontal periphery of the screen. For an anti-saccade trial (cued by red cross-hair), subjects are instructed to look at the screen location directly opposite from the location of the dot. Both the pro-saccade and anti-saccade blocks consist of 30 seconds of continuous trials of the corresponding type. Thus the anti-saccade *task* compares two *conditions* – anti-saccade vs. pro-saccade. By comparing the signal levels for these two conditions, it is possible to identify the cortical regions that are activated during inhibition of reflexive eye movements.



**Figure 16.** Difference between pro-saccade and anti-saccade conditions. a) During the pro-saccades condition (cued by green cross), the subject is instructed to look towards the dot when it appears. b) During the anti-saccades condition, the subject is instructed to look away from the location of the dot (inhibit reflexive pro-saccadic eye movement). In both cases, the probe randomly alternates between Probe 1 and Probe 2 (see Figure 15).

### 3.1.2 Data collection protocol

A 3.0 Tesla General Electric Signa Scanner is used for the study. Three sets of structural scans are collected prior to obtaining the functional data – these anatomical images are used for spatial normalization to a common coordinate system for group analysis. The scanning parameters are TR = 2500 msec, TE = 18 msec, flip angle = 70 degrees, slice thickness = 3.2 mm with 0 mm gaps, number of slices = 30, field of view (FOV) = 20 cm, number of excitations (NEX) = 1. The VGS and anti-saccade tasks are each 6:00 minutes in length. The functional scan consists of continuous imaging of the brain during the eye-movement tasks – yielding a total of 144 image volumes per run for each task.

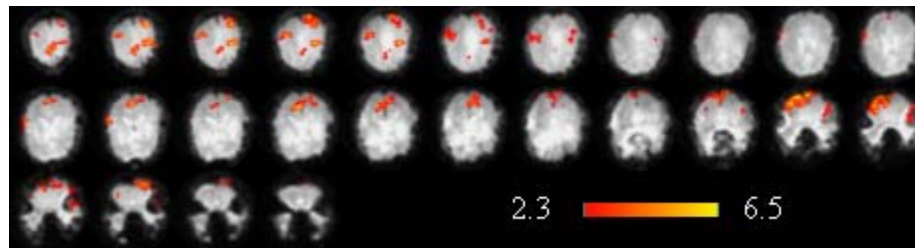
## 3.2 DATA ANALYSIS

### 3.2.1 Pre-processing

First, the structural and functional scans are re-constructed from k-space to 3D image volumes (see Chapter 2). Next, each of the functional images is corrected for head-motion by a rigid-body transformation to the mean of the functional images. Also, the time-series for each voxel is corrected for spikes and linear trends. Finally, the functional images are spatially smoothed with a 3D Gaussian kernel (FWHM=4mm) prior to statistical analysis.

### 3.2.2 Statistical Analysis of Single Brain

For each voxel, a GLM-based statistical model (Chapter 2) is employed to determine the degree to which the pre-processed time-series for the voxel matches the box-car pattern of the task timecourse. The score from the statistical test is a measure of the level of neural ‘activation’ associated with performance of a particular task (e.g. saccadic eye movements). These statistical scores are thresholded and the voxels above the threshold are mapped to a 3D image of the brain to display the regions of ‘significant’ activation. Axial slices of a 3D activation image are shown in Figure 17. As can be noted in the figure, the activation is concentrated within a few regions of activation (ROA).



**Figure 17.** Example activation image for the VGS task (axial slices of 3D image volume). For each voxel, the z-score is represented as a heat-map.

### 3.2.3 Group Analysis

Once activation images are created for individual subjects, they are analyzed for the presence of patterns that can be used to discriminate between the two groups. For example, hypothetical differences in sizes of regions of activation are illustrated in Figure 18.



**Figure 18.** Hypothetical differences in activation patterns between two populations of subjects.

The goal of the knowledge discovery process is to automatically report the presence of such differences between groups. While the difference between the two groups in this illustrative example is readily apparent from visual inspection, the problem becomes more difficult with many regions of activation in a 3D image. Also, the differences in shapes and sizes of individual brains add to difficulty of the problem – the images must be ‘spatially normalized’ to a common coordinate-system before individual voxels can be compared. As discussed earlier, current methods of spatial normalization cannot fully compensate for differences between individual brains. For this reason, this work explores an alternative to voxel-based knowledge discovery methods that does not require perfect spatial normalization of the different brains.

## 4.0 KDSF FRAMEWORK

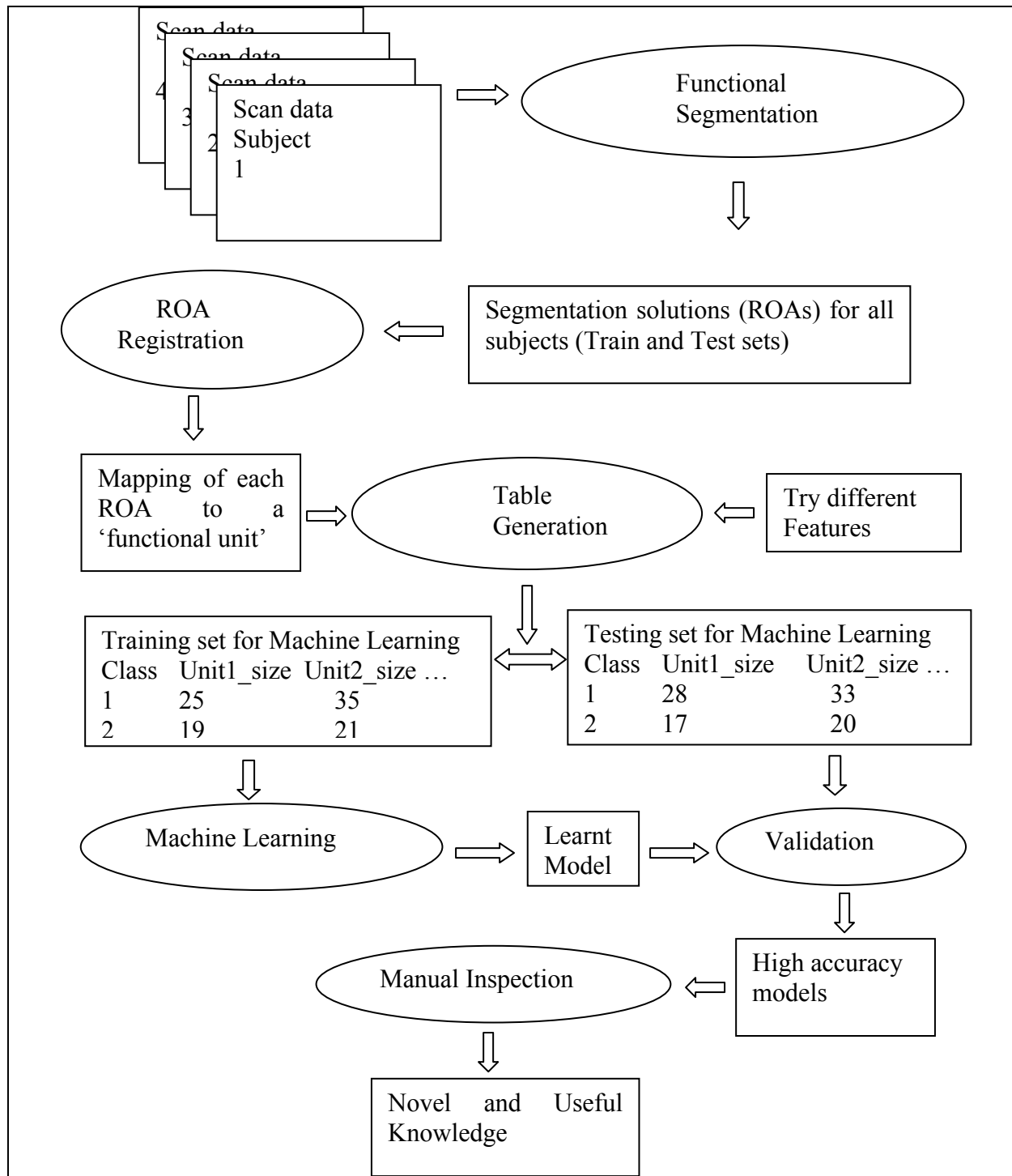
Knowledge Discovery from Segmented fMRI (KDSf) is a proposed framework for automated knowledge discovery from fMRI image datasets. Given a set of labeled fMRI images (e.g. labeled as ‘normal’/‘disease’), the framework reports classification models that can accurately discriminate between the two groups. The discovered knowledge may be stated explicitly in the classification model (as in ‘if-then-rules’) or it may be implicit (as in Support Vector Machines). In the latter case, the feature set used by the model is manually inspected to gain insight into the relationship exploited by the classification model.

The KDSf framework is an alternative to the conventional knowledge discovery process based upon statistical hypothesis testing. As discussed earlier, knowledge based upon classification models is more amenable to clinical applications than the traditional approach of reporting p-values for regions of the brain that exhibit differential activation.

There are four steps in the KDSf framework:

1. A functional segmentation step to identify the Regions of Activation (ROAs) in the images.
2. A ROA registration step to relate ROAs from different subjects.
3. A feature construction step that creates data tables for machine learning.
4. Automated knowledge discovery.

The steps of the KDSf framework are shown in Figure 19. Note that the framework is not tied to specific algorithms – while algorithms are provided in this work for these steps, it is possible to substitute other algorithms. The four steps are explained in the following sections.

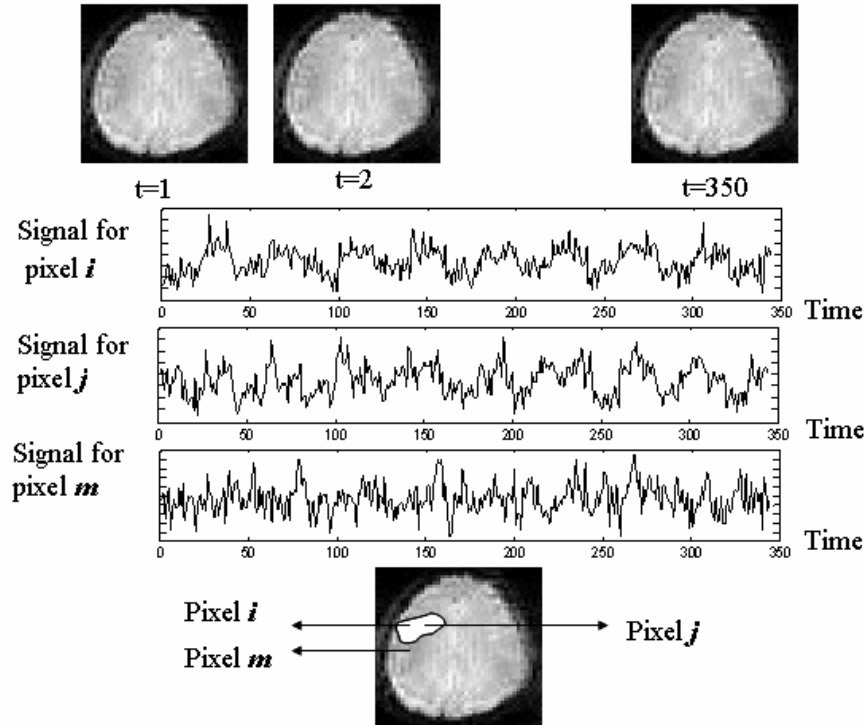


**Figure 19.** Outline of the KDSf framework for Knowledge Discovery.

## 4.1 DESCRIPTION OF KDSF

### 4.1.1 Functional Segmentation

The goal of the functional segmentation step is to identify regions of activation (ROA) in the fMRI data from a single subject. Functional segmentation can be performed in one of several ways. For example, it is possible to create a set of ROAs by thresholding a conventional activation map (SPM) for the subject (see Figure 17). The drawback of this approach is that the sizes of the segments are dependent upon the threshold – the threshold does not reflect any natural boundary in the image. Also, dissimilar time-courses with similar t-scores can be incorporated within a functional segment. Another possibility is to use standard image processing techniques to segment the activation image – however the diffuse intensity transitions in typical activation images are not well suited for intensity-based segmentation. Alternatively, the images can be segmented based upon similarity of time-courses. For example, ICA can be employed to identify a set of spatial components from a 4D image – the component whose timecourse is most similar to the stimulus timecourse can be thresholded to segment the image. However, note that the ICA method does not use spatial relationships between timecourses for isolation of the components – the image is ‘flattened’ prior to analysis. Another approach to functional segmentation can be based upon similarity of timecourses from *spatially adjacent* voxels – this is the approach taken in this work. A schematic of this approach to the functional segmentation problem is shown in Figure 20.



**Figure 20.** Functional segmentation based upon similarity of adjacent time-courses. The top row represents the time-series of functional images. Timecourses for pixels within the Region of Activation (e.g. pixels  $i$  and  $j$  in bottom image) are similar compared to the timecourses outside the ROA (e.g. pixel  $m$ ).

While functional segmentation based upon similarity of timecourses has the ability to detect natural boundaries in the patterns of time-courses, the method can be sensitive to the presence of confounding artifacts in the time-courses (e.g. from un-corrected head-motion of subject, see Chapter 5). Thus functional segmentation may include both ‘activated’ segments and segments which reflect some confounding process such as head-motion. It is necessary to remove the latter type of segments prior to machine learning. This filtering of segments is achieved by comparison of the segment timecourses with the stimulus timecourse – standard methods (t-test, GLM, correlation etc.) can be used for this purpose.

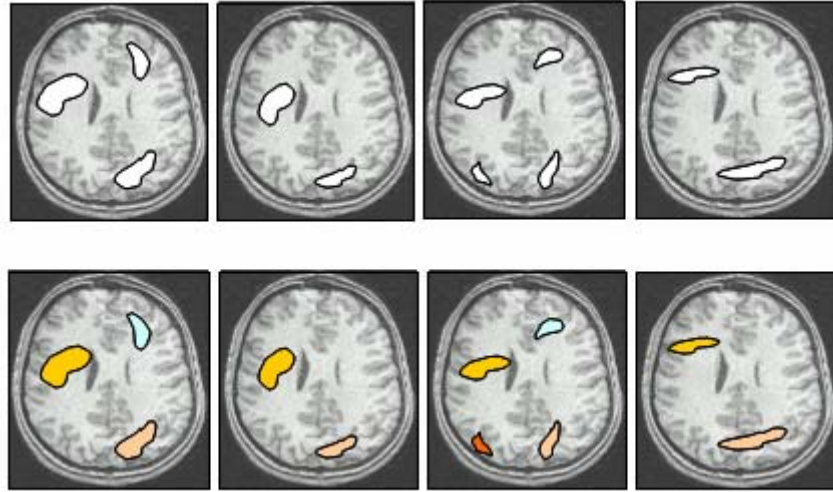
Regardless of the method employed for functional segmentation, functional segmentation is essentially a data reduction technique that represents the activation information from an fMRI scan of a subject with a handful of ROAs. The characteristics of these ROAs are used as features for machine learning.

### 4.1.2 ROA Registration

The KDSf framework is proposed for knowledge discovery in the absence of precise normalization of the morphological differences between individual brains. After the ROAs are identified by functional segmentation, it is necessary to match ROAs across brains of different subjects. Due to anatomical and functional variation between subjects, and the imprecise nature of the current techniques for spatial normalization, it is unrealistic to expect ROAs from different subjects to be perfectly aligned with each other, even after spatial normalization. Thus, it is necessary to determine whether two spatially normalized ROAs from two different subjects are possibly instances of the same ‘functional unit’. For the purposes of this framework, it is assumed that a concentration of ROAs from different subjects in approximately the same anatomical location reflects a ‘functional unit’ of the brain. For example, Broca’s area is an example of a ‘functional unit’ that is involved in language processing, speech production and comprehension. Other examples of functional units are illustrated in Chapter 3. The fMRI paradigm may activate several functional units in individual brains. Also, some functional units may not be represented in each brain (see Figure 21).

Knowledge discovery in terms of characteristics of functional units requires that ‘nearby’ ROAs in different brains be labeled as instances of the same functional unit – this is the goal of the ROA registration step. Note that this is a data driven process, it is not necessary to pre-define the approximate locations of the functional units. The ROA registration step is illustrated in Figure 21.



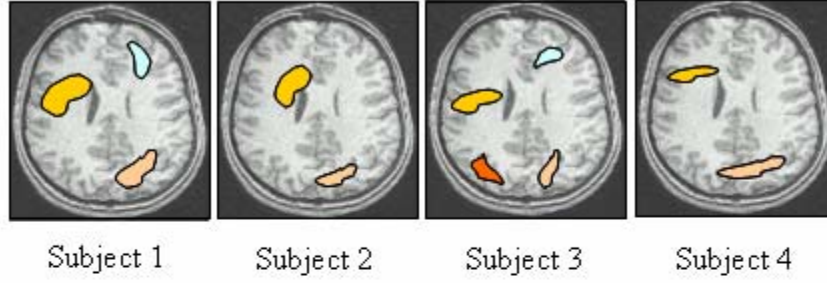






**Figure 21.** ROA registration determines the correspondence of ROAs across subjects. The ROAs from a set of segmented images (top row) are labeled (colored) by the registration process (bottom row) – each color corresponds to a functional unit.

ROA registration is a challenging problem and it is difficult to find optimality criteria that does not require anatomical heuristics. Instead of proposing new anatomical heuristics, a simple registration method – Greedy ROA Registration (GRR) is presented here. This method relies upon a user-specified distance parameter that controls the maximum displacement between ROAs that are instances of the same functional unit – anatomical constraints (e.g. lobar boundaries) are not enforced. While this may be too simplistic for real-world use, it is sufficient for demonstration of the concept. The details of the GRR method are presented in the Appendix.

#### 4.1.3 Feature Construction

After each ROA is mapped to a functional unit, features are constructed for each functional unit. For a functional unit, these feature-values may be the sizes of the corresponding ROAs, average activation strength within the ROAs, or other characteristics of the ROAs (e.g. centroid or shape). Thus, the columns of the machine learning table are the features constructed from functional units, and the rows represent individual brains. Note that some brains may not exhibit an ROA corresponding to some functional unit – these functional units can be dropped from the data table in case of insufficient representation in the population. Figure 22 illustrates the feature construction step from a set of images after region registration.



Subject	Class	Size	t-score	Size	t-score
		Unit1 	Unit1 	Unit2 	Unit2 
1	0	24	5.01	14	4.23
2	0	22	4.81	10	4.78
3	1	16	6.03	11	5.76
4	1	10	5.89	16	4.73

**Figure 22.** KDSf feature construction: The machine learning table is constructed from attributes of functional units (e.g. size and average t-score). The functional units are labeled with colors.

#### 4.1.4 Automated Knowledge Discovery

Once features are constructed from functional units, the knowledge discovery step involves a search for combinations of features that result in accurate classification models. For example, to detect univariate size differences for a functional unit between groups, a single feature machine learning table can be constructed for each functional unit. Alternatively, to detect multivariate differences, all size-based features for all the functional units can be included in the machine learning table.

Note that the region registration step in KDSf requires that both the training set images and testing set images are registered together to obtain a common set of functional units for comparison. Even though the testing set is utilized during feature construction, this feature construction step is ‘unsupervised’ in the sense that the class-labels are not used during feature construction. Unsupervised feature construction from the full dataset (training plus testing) is also employed by Principal Component Analysis (PCA).

Once a particular machine learning table is created for a particular feature set, different machine learning methods (e.g. Neural networks, SVM etc.) are employed to identify the best

possible model for discrimination based upon the selected feature set. To guard against over-fitting, the classification models are cross-validated with testing sets that were not used for training the classification model. The creation of random training and testing sets can be repeated a substantial number of times to get additional confidence in the classification accuracy of the selected feature set. Those models with high classification accuracies are reported (along with the feature set) as potentially ‘interesting’. These high accuracy models are manually inspected to determine the validity (and novelty) of the knowledge embedded in the models.

Note that this approach to automated knowledge discovery is not restricted to features constructed from functional units – other feature construction methods such as KBV and PCA can also be employed. It is hypothesized that the segmented approach (KDSf) can be more effective than knowledge discovery with KBV or PCA. The effectiveness of this approach to automated knowledge discovery is critically dependent upon classification accuracies – if the classification methods are unable to detect inter-group differences that are actually present in the set of images, knowledge discovery is not possible. Thus, to validate the hypothesis regarding the effectiveness of KDSf, it is necessary to compare classification accuracies for different feature construction methods.

## **4.2 EVALUATION OF KDSF**

The KDSf framework is a high-level description of a novel approach to automated knowledge discovery from fMRI datasets. The effectiveness of the framework for knowledge discovery from real-world datasets is dependent upon the actual algorithms employed by the framework and the quality of the fMRI dataset. For this reason, the empirical effectiveness of the framework under controlled conditions is evaluated here with synthetic data, which facilitates systematic variation of the parameters of interest. The goal of this evaluation is to determine the set of circumstances under which the KDSf approach to feature construction has an empirical advantage over other methods of feature construction (PCA and KBV) – actual results with real-world datasets will vary.

In this section, the validity of the claim regarding improved classification accuracies with KDSf is assessed with systematic variation of the inter-group differences of interest. Two kinds

of differences are assessed in this study – differences in activation strengths between groups, and differences in sizes of activated regions between groups. These two cases are studied in detail since these situations are expected to be frequently encountered in actual fMRI datasets. Also, in the presence of imprecise spatial normalization, the conventional voxel-based methods may be inadequate in these cases. However, it should be noted that voxel-based approaches can be superior if the difference between populations is restricted to the shapes of regions of activations (this situation is probably rare in practice). Also, if different locations in the brain are activated for different groups, the voxel-based methods (with suitable smoothing) may be equally effective in detecting this difference. Thus, KDSf is not a replacement for conventional methods of feature construction, KDSf complements PCA and KBV.

While the KDSf framework is applicable to both 2D and 3D brain images, the formal evaluation in this work is based upon a publicly available 2D fMRI dataset that has been used for other fMRI evaluation studies [51]. Since large numbers of synthetic datasets are processed for systematic evaluation, the lower computational cost for 2D images was also a factor for adopting this approach. Also, to avoid introduction of irrelevant features for KBV and PCA, only one simulated region of activation (ROA) was introduced in the images. This is because, even if only one ROA exhibits between-group differences, the presence of multiple ROAs complicates the feature construction task for KBV and PCA – however this is not a problem for KDSf if the ROAs are well separated from each other. The consideration of only one ROA simplifies the comparison of accuracies for the different methods of feature construction. Thus, since only one ROA is present in the simulated activation images, region registration is not required for these simulated datasets.

#### **4.2.1 Baseline Data**

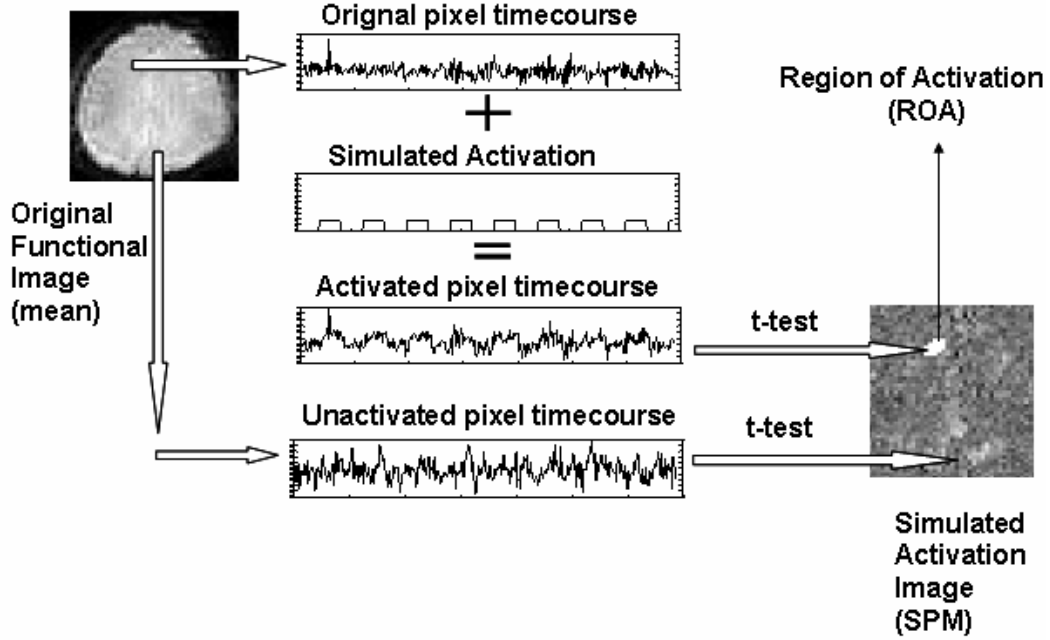
The synthetic datasets were based upon a publicly available benchmark dataset used for comparison of various clustering techniques for functional segmentation [51]. The baseline in vivo MRI dataset consisted of time-series of 334 images with a matrix size of 64x64 pixels per image (one 2D axial slice of the brain). Since the subject was scanned ‘at rest’, no task-specific activation was expected to be present in the baseline data. To simulate task-related activation, the timecourses for the pixels in a region of the brain were modulated by a box-car activation pattern

(20 on, 20 off, repeated 8 times). Since in vivo noise is known to be temporally auto-correlated, this hybrid approach of mixing simulated activation with actual time-series from scans is more realistic than datasets from mathematical ‘phantoms’.

Evaluation of KDSf required a number of synthetic activation images (120 simulated subjects per dataset). This set of synthetic activation images were constructed from the original in vivo dataset as follows. First, a baseline activation image was constructed by performing a two-sample t-test (comparing signal values during the ‘on’ and ‘off’ conditions) on the original (un-activated) timecourse from each pixel in the image. To introduce inter-subject variation in the baseline activation image, Gaussian noise was added to this baseline activation image (zero mean, and standard deviation equaling 0.05% of the computed standard deviation of the baseline activation image – this level of inter-subject variation was chosen to avoid drastic changes to the overall baseline activation pattern). Next, for each subject, task-induced ‘activation’ was introduced to some of the pixels of the baseline activation image – this is described next.

For each subject, the characteristics of the region of simulated activation were chosen by sampling from a probability distribution. These characteristics included the strength of activation (as measured by CNR), the shape, and the size of the region of activation. For each pixel inside the region of activation, the original time-course was modulated by the on-off activation pattern. The strength of the activation (increase in signal value) was computed from the CNR desired for the particular subject (e.g. based upon ‘normal’/‘disease’ label). As in the original study [51] for this dataset, an average noise level (94.5 in MR units) inside the brain was used for CNR calculations. However, to introduce small differences in activation levels between pixels (across simulated subjects), Gaussian noise (with zero mean and standard deviation of 0.05% of 94.5) was added to the noise level prior to computation of the activation strength for individual pixels. The on-off activation pattern (scaled to the specified CNR, using the noise level for the pixel) was added to the original time-course to simulate activation. Finally, the activation value for the task-modulated pixel was computed by a t-test on the modified time-course.

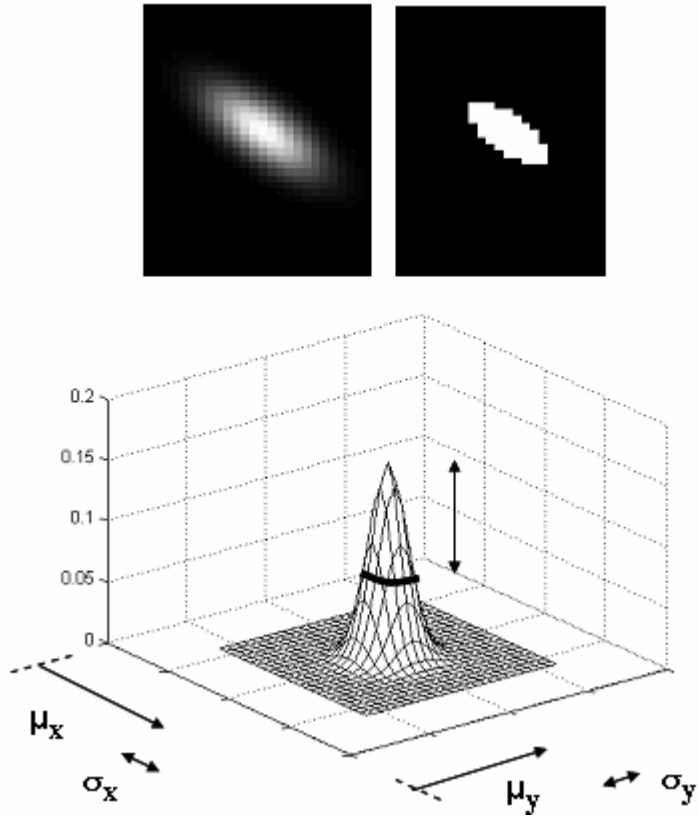
Thus, starting with the baseline activation image (with no task-induced activation), the addition of simulated activation to all the pixels inside the region of activation yielded a complete activation image for one subject. This process is illustrated by Figure 23.



**Figure 23.** Schematic for creation of synthetic activation images. Simulated activation (on-off pattern) is added to the time-series for the set of pixels inside the region of activation (top arrow).

#### 4.2.2 Model of Region of Activation (ROA)

The inter-subject differences in the characteristics of the region of activation (arising from anatomical/functional variation and imperfect spatial normalization) are modeled by a probability distribution. The spatial boundary (location and shape) of the region of activation is modeled by the ‘shape’ of a bivariate Gaussian function, truncated at half the maximum value of the Gaussian (see Figure 24). The parameters for this 2D Gaussian distribution ( $\mu_x$ ,  $\mu_y$ ,  $\sigma_x$ ,  $\sigma_y$ ,  $\rho_{xy}$ ) are sampled from independent Gaussian distributions which are part of a generative model (described below). The average activation level (measured by CNR) inside the region of activation is also sampled from a Gaussian distribution in the generative model. Note that all the pixels in the simulated region of activation share similar activation strengths (with Gaussian noise as described above). The activation level does not decay outward as implied by the 2D Gaussian – this convention is adopted from the original study [51] that employed this dataset.



**Figure 24.** A randomly generated 2D Gaussian function is thresholded to generate random shapes for regions of activation. The pixels in the 2D Gaussian (top, left) for which the value exceeds half of the maximum value (vertical arrow) are retained in the region of activation (top, right).

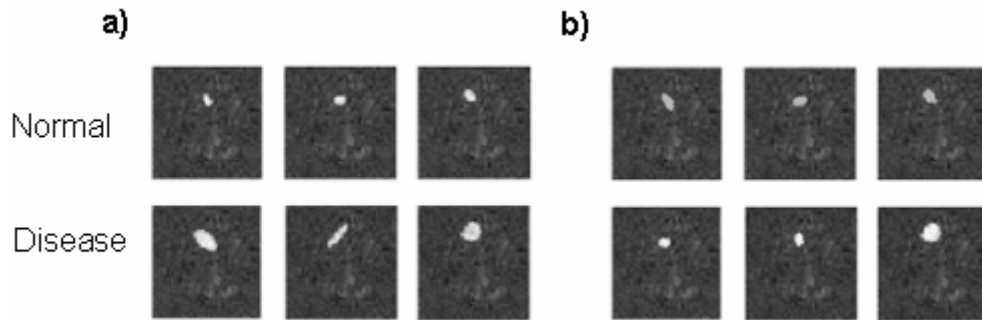
**Table 3.** Parameters for simulated regions of activation (ROA).

ROA Parameter	Explanation
CNR	Average activation level in region (relative to average noise)
$\mu_x$	Centroid of activated region along x-axis
$\mu_y$	Centroid of activated region along y-axis
$\sigma_x$	Measure of ‘span’ of activated region along x-axis
$\sigma_y$	Measure of ‘span’ of activated region along y-axis
$\rho_{xy}$	‘Eccentricity’ of shape of activated region (coefficient of correlation of the bivariate Gaussian distribution)

### 4.2.3 Generative Model

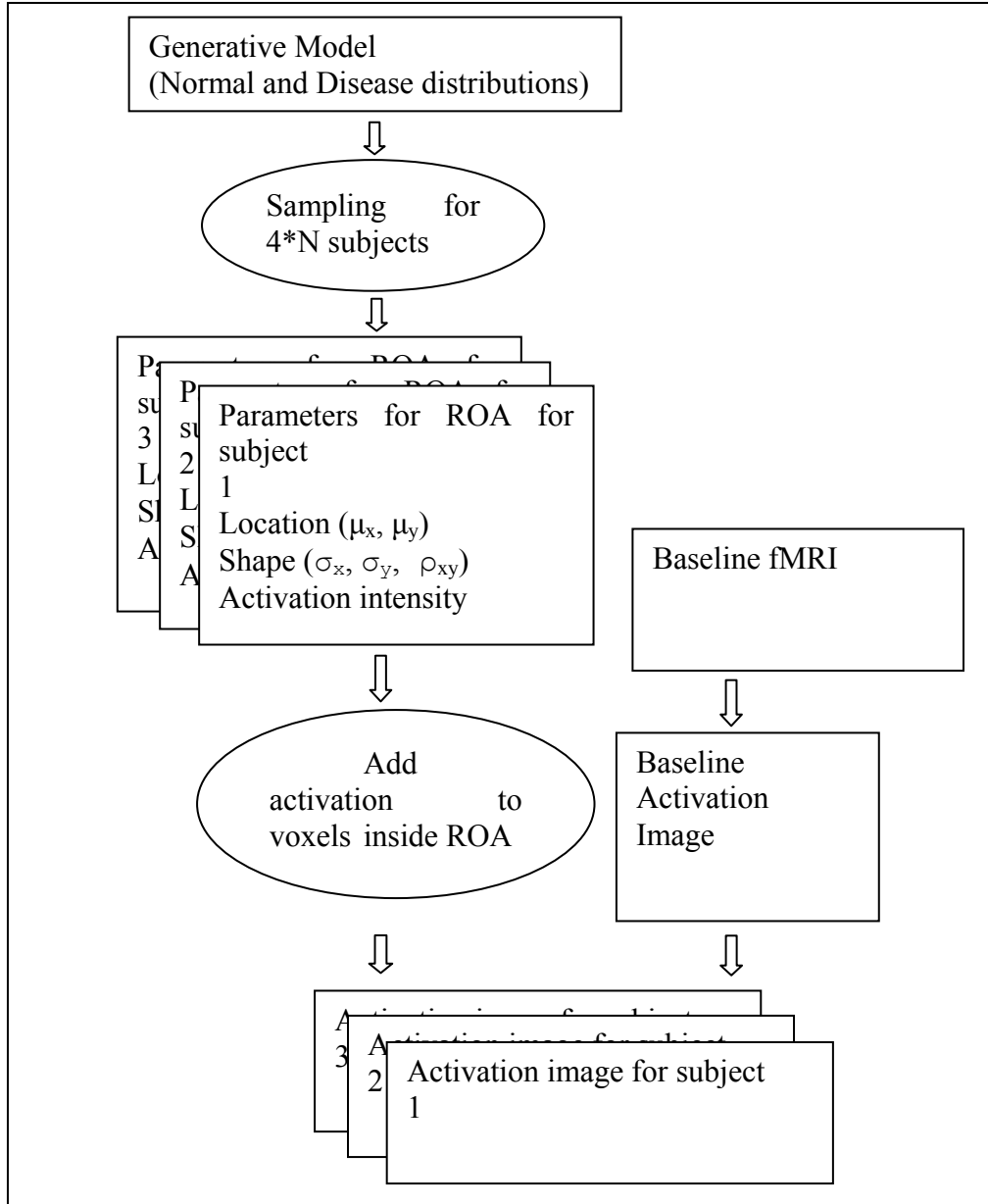
The probability distributions for the parameters of the single region of activation (Table 3) are specified by a generative model. Thus a generative model is a mechanism for creating a set of activation images with a shared underlying distribution for the region of activation (ROA). A generative model consists of a set of independent Gaussian distributions which control the ROA parameters for the dataset (for both ‘normal’ and ‘disease’ images).

The generative model specifies different Gaussian distributions for the two groups (‘normal’/‘disease’) to simulate differences between populations. For example, the activated region could exhibit a larger mean size (and/or activation level) in the disease population compared to the normal population. The goal of the machine learning techniques is to detect these differences automatically, in the presence of inter-subject variability of location of the ROA (or, location-variability in short). The location-variability is modeled by probability distributions for  $\mu_x$  and  $\mu_y$  – larger standard deviations for these two ROA parameters reflect higher anatomical/functional variability in the population. A set of example activation images is shown in Figure 25. The steps for the generation of a dataset of synthetic activation images incorporating inter-group differences is shown in Figure 26.



**Figure 25.** Example simulated activation images from datasets created with generative models. a) Between-group differences in ROA-size. b) Between-group differences in activation-level (indicated by brightness of the ROA). In both cases, ROAs exhibit between-subject location-variability and size-variability.





**Figure 26.** Steps for generation of synthetic dataset. Multiple activation images are created from a generative model.

The parameters for a generative model (covering both ‘normal’ and ‘disease’ subjects) are as follows:

- $N$ , the number of subjects in each group in training/testing set
- $sd(\mu)$ , the variability of location of the ROA in images (same for both dimensions of image, but  $\mu_x$  and  $\mu_y$  are sampled independently)

$$sd(\mu) = sd(\mu_x) = sd(\mu_y) \quad (4.2.1)$$

- $\sigma^N$ , the ‘span’ of the ROA for ‘normal’ subjects (same for both dimensions of image, but  $\sigma_x$  and  $\sigma_y$  are sampled independently)

$$\sigma^N = \text{mean}(\sigma_x) = \text{mean}(\sigma_y) \quad (4.2.2)$$

- $\sigma^{D/N}$ , the ratio of average ROA span for disease subjects to that for ‘normal’ subjects ( $\sigma^D$  refers to the average ROA span for ‘disease’ subjects and  $\sigma^N$  refers to the average ROA span for ‘normal’ subjects)

$$\sigma^{D/N} = \sigma^D / \sigma^N \quad (4.2.3)$$

- $\text{sd}(\sigma^N)$ , the between-subject variability of the span of ROAs (same for both ‘normal’ and ‘disease’ subjects)

$$\text{sd}(\sigma^N) = \text{sd}(\sigma^D) \quad (4.2.4)$$

- $\text{CNR}^N$ , the average activation strength for ROAs in ‘normal’ subjects
- $\text{CNR}^{D/N}$ , the ratio of average activation strength for ‘disease’ subjects to that for ‘normal’ subjects ( $\text{CNR}^D$  is the average activation strength for ROAs in ‘disease’ subjects)

$$\text{CNR}^{D/N} = \text{CNR}^D / \text{CNR}^N \quad (4.2.5)$$

- $\text{sd}(\text{CNR}^N)$ , the variability of activation strengths between subjects (same in both groups)

$$\text{sd}(\text{CNR}^N) = \text{sd}(\text{CNR}^D) \quad (4.2.6)$$

The parameters and their units are summarized in Table 4.

**Table 4.** Parameters for generative model for a population of activation images.

<b>Generative Model Parameter</b>	<b>Explanation</b>	<b>Units</b>
N	Number of subjects representing each group in training/testing set	Subjects
sd( $\mu$ )	The variability of location of the ROA	Pixels
$\sigma^N$	The ‘span’ of the ROA for ‘normal’ subjects	Pixels
$\sigma^{D/N}$	Ratio of mean ‘span’ of ROAs in disease group to that in normal group	None (ratio)
sd( $\sigma^N$ )	The between-subject variability of ‘span’ of ROAs	Percentage of $\sigma^N$
CNR <sup>N</sup>	The mean activation strength for ‘normal’ subjects	None (ratio of activation level to noise level)
CNR <sup>D/N</sup>	Ratio of mean CNR in disease group to that in normal group	None (ratio)
sd(CNR <sup>N</sup> )	Variability of CNR between subjects in both groups	Percentage of CNR <sup>N</sup>

The relationships between the generative model parameters and the parameters for the ROAs are illustrated in Table 5. Some of the generative model parameters are not varied for this evaluation – these fixed values are also shown in Table 5.

**Table 5.** Relationship between generative model parameters and ROA parameters. The Gaussian distributions for ROA parameters (rows) are controlled by the parameters of the generative model. Note ‘\*’ signifies multiplication.

<b>ROA Parameter</b>	<b>Mean for ‘normal’ group</b>	<b>Standard deviation for ‘normal’ group</b>	<b>Mean for ‘disease’ group</b>	<b>Standard deviation for ‘disease’ group</b>
CNR	CNR <sup>N</sup>	sd( CNR <sup>N</sup> )	CNR <sup>N*</sup> CNR <sup>D/N</sup>	sd( CNR <sup>N</sup> )
$\mu_x$	30	sd( $\mu$ )	30	sd( $\mu$ )
$\mu_y$	22	sd( $\mu$ )	22	sd( $\mu$ )
$\sigma_x$	$\sigma^N=2.5$	sd( $\sigma^N$ )	$2.5*\sigma^{D/N}$	sd( $\sigma^N$ )
$\sigma_y$	$\sigma^N=2.5$	sd( $\sigma^N$ )	$2.5*\sigma^{D/N}$	sd( $\sigma^N$ )
$\rho_{xy}$	0	0.3	0	0.3

#### 4.2.4 Research Design

To test the hypothesis regarding improved classification accuracy of the KDSf framework compared to conventional methods of feature construction (KBV and PCA), the parameters of the generative model (above) are systematically varied to simulate a wide variety of conditions. The validity of the hypothesis is evaluated for the following simulated situations:

1. The activated regions in the ‘disease’ population are larger than those in the ‘normal’ population.
2. The activated regions in the ‘disease’ population exhibit higher activation levels, compared to the ‘normal’ population.

Note that, by switching the two class labels (‘disease’ and ‘normal’), the results from these two situations also apply to the reverse situations where the activated regions are smaller (or less activated) in the ‘disease’ population.

For sensitivity analysis of the validity of the hypothesis for these two situations, the parameters of the generative model are varied such that the resulting datasets reflect different degrees of variations in size and activation levels. For test situation 1 (referred to as Test-Suite-Size), various degrees of between-group differences in size are studied. Sizes of ROAs in the disease group are controlled by  $\sigma^{D/N}$  – the ratio of the mean ROA ‘span’ in the disease group compared to that in the normal group. In Test-Suite-Size, the average ROA span in the disease group was 10% to 100% higher than that in the normal group.

Similarly, for test situation 2 (referred to as Test-Suite-CNR), between-group differences in activation levels ranging from 10% to 100% are studied. Table 6 shows the combinations of generating model parameters that are studied for the two situations.

Test-Suite-Size considers whether between-group size differences for the ROA can be detected by conventional feature construction methods in the presence of inter-subject variability in the locations of activation (location-variability). Thus, in Test-Suite-Size, three characteristics of ROAs were explored – between-group ‘size’ differences, between-subject location-variability, and between-subject size-variability. The effect of spatial smoothing is also studied.

The second test situation (Test-Suite-CNR) considers whether conventional feature construction methods (along with spatial smoothing) can detect between-group differences in activation levels in the presence of location-variability – in other words, can smoothing

compensate for imprecise spatial normalization. Thus, for Test-Suite-CNR, between-group differences in activation and between-subject location-variability are studied in conjunction with different levels of spatial smoothing.

Each combination of values for the generative model parameters (Table 6) is a distinct generative model that yields a dataset with simulated differences between the normal and disease populations. For Test-Suite-Size, 80 different generative models are studied – for each generative model 12 datasets are generated. Each dataset consists of training and testing sets with equal numbers of normal and disease subjects in each set. The 12 repetitions for each generative model provide a mechanism for testing the statistical significance of the observed differences in classification accuracy.

**Table 6.** Generative model parameters explored in Test-Suite-Size and Test-Suite-CNR.

<b>Generative Model Parameter</b>	<b>Values studied for Test-Suite-Size</b>	<b>Values studied for Test-Suite-CNR</b>	<b>Description (units)</b>
N	30	30	Number of subjects in each class in training/testing sets (subjects)
sd( $\mu$ )	0, 1, 2, 4	0, 1, 2, 4	Between-subject location variability of ROA (pixels)
$\sigma^N$	2.5	2.5	‘Span’ of ROA (pixels)
$\sigma^{D/N}$	1, 1.1, 1.2, 1.5, 2	1	Between-group ‘size’ variability (none)
sd( $\sigma^N$ )	0, 0.1, 0.2, 0.3	0.1	Between-subject ROA ‘size’ variability (Percentage of $\sigma^N$ )
CNR <sup>N</sup>	1.66	1.66	Ratio of activation level to noise level in ‘normal’ group (none)
CNR <sup>D/N</sup>	1	1, 1.1, 1.2, 1.5, 2	Between-group activation differences (ratio)
sd(CNR <sup>N</sup> )	0	0.05	Between-subject variability of activation levels (percentage of CNR <sup>N</sup> )

#### 4.2.5 Feature Construction/Selection

For each image dataset, features are constructed using all three methods of interest (KDSf, PCA and KBV). Since the hypothesis is about classification accuracies for different feature construction methods (rather than about differences between machine learning methods) the best classification accuracy from three different machine learning methods (GNB, ANN and SVM) are used for comparison between the three feature construction methods. Also, for KBV and PCA, the effect of feature selection was considered by determining the best accuracy over the parameter-space for these methods ( $k$  in KBV and number of components in PCA).

For voxel-based feature construction, the voxels (pixels in this case) were first sorted by an interestingness measure (equation 2.4.5) – only the training set images were considered for this step. For  $k$ -best-voxels (KBV) feature construction, where  $k$  ‘best’ (or most interesting) voxels are retained in the machine learning model,  $k$  was varied from 1 to the total number of pixels in the image ( $M$ ). However, since  $M$  is large, an un-even step-size was used to cover the range from 1 to  $M$ . All values from  $k=1$  to  $k=100$  were used; beyond 100, the  $k$  was incremented by 500 at each step till  $M$  was reached. Since the average size of the ROA is around 25 pixels for the normal group, the best accuracy is generally reached for  $k$  less than 100.

For PCA-based feature construction, all the images in the dataset (both training and testing set images) were processed together to compute the principal components (eigenimages). For feature selection, the eigenimages were included in the model in the order of the eigenvalues [2] (i.e. ordered by importance of the component in explaining the variance of the data). Thus, the number of eigenimages in the selected feature-set was varied from 1 to the total number of eigenimages ( $4*N$ , where  $N$  is the number of subjects in each class for training/testing datasets).

For KDSf, features were constructed by functional segmentation of the activation images (the ROA boundaries were known for synthetic images). Since only one ROA was modeled by the generative models, region registration was not required. The feature selection step consisted of choosing the relevant attribute of the ROA – the number of pixels in the ROA for Test-Suite-Size, and the average activation level of the ROA (computed by averaging the t-scores inside the ROA) for Test-Suite-CNR. Note that for the purposes of this evaluation, it is assumed that some method of functional segmentation (with 100% accuracy) is available. The accuracy of the specific functional segmentation method (ACEIC) is evaluated separately (Chapter 5).

#### 4.2.6 Hypothesis Testing

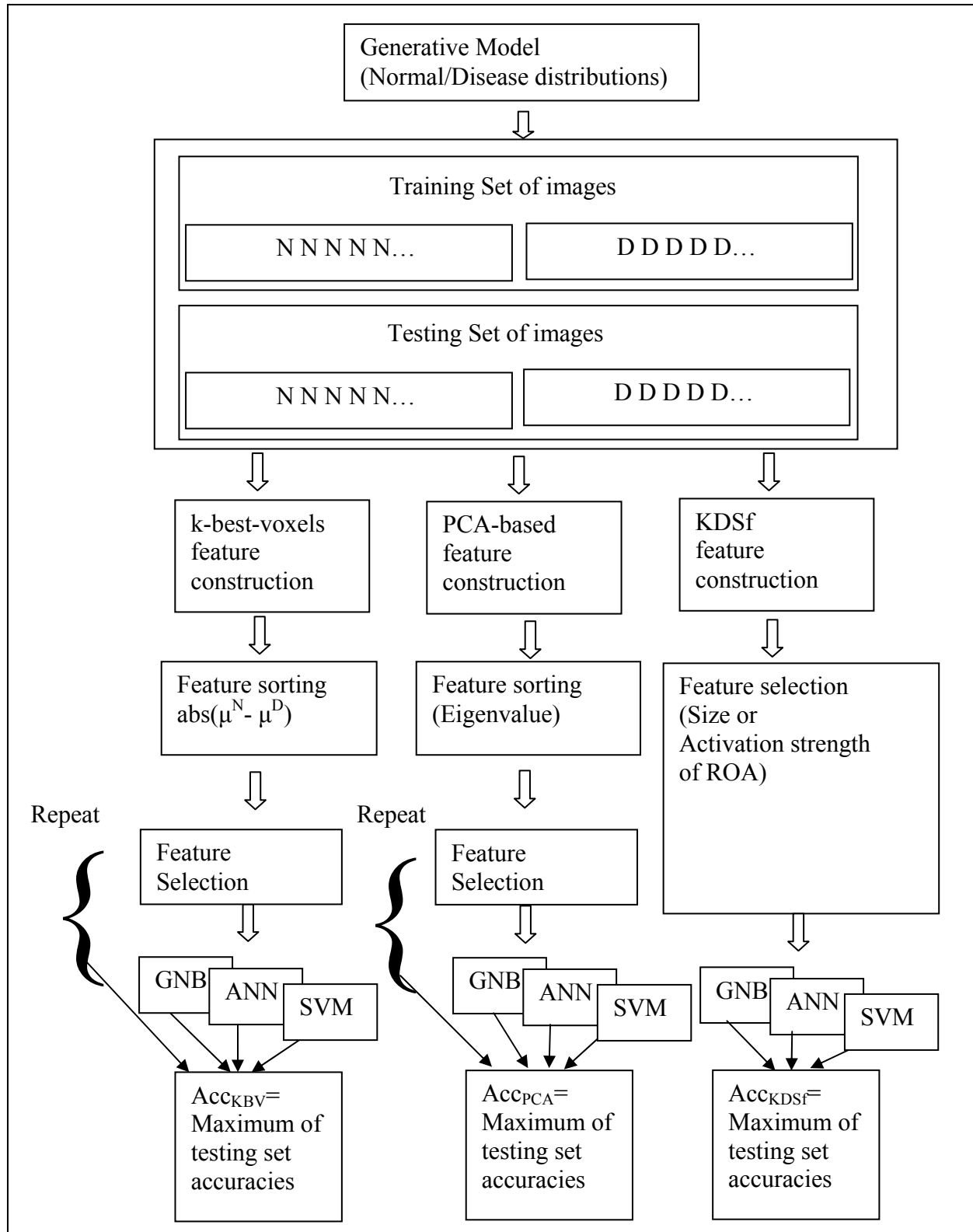
The test situations described above (Test-Suite-Size and Test-Suite-CNR) are designed to test the hypothesis regarding higher classification accuracy of KDSf compared to PCA or KBV. For testing the statistical significance of the observed differences in accuracy, 12 image datasets were generated for each generative model. The 12 classification accuracies from each of the feature construction methods (twelve instances of  $Acc_{KBV}$ ,  $Acc_{PCA}$ ,  $Acc_{KDSf}$  in Figure 27) provide a basis for testing the statistical significance of the observed differences in accuracy between methods. For a given generative model, the following sub-hypotheses are tested:

$$\text{Mean}(Acc_{KDSf}) > \text{Mean}(Acc_{KBV}) \quad (\text{Hypothesis 4.1})$$

and

$$\text{Mean}(Acc_{KDSf}) > \text{Mean}(Acc_{PCA}) \quad (\text{Hypothesis 4.2})$$

Since the same dataset is used to compute the accuracies for the three cases (KDSf, PCA, KBV), paired t-tests were used for hypothesis testing. For a given generative model, if these sub-hypotheses were accepted (at  $\alpha=0.05$ , with Bonferroni correction), then the overall hypothesis was accepted for the particular generative model.



**Figure 27.** Overview of accuracy computations (for KDSf, PCA and KBV) from a synthetic dataset created with a generative model.



#### 4.2.7 KDSf and segmentation accuracy

Apart from testing the hypothesis regarding improved classification accuracy of KDSf, the dependence of the accuracy of KDSf on the accuracy of the functional segmentation was also explored. Simulation studies were performed to explicitly demonstrate this dependence of KDSf accuracy on segmentation accuracy. Thus, the accuracy improvements achieved by the ACEIC method of functional segmentation (Chapter 5) has a direct relevance to knowledge discovery from image datasets.

For this demonstration with datasets from Test-Suite-Size, classification accuracy with inaccurate segmentation is compared with classification accuracy with accurate segmentation. To simulate inaccurate segmentation, two operators are defined for a region of activation. Give a region of activation, the ‘erosion’ operator removes the pixels in the ‘internal boundary’ (see section 5.2) and the ‘dilation’ operator adds the pixels in the ‘external boundary’ (see section 5.2). Note that these operators are somewhat different than the mathematical morphology operations of the same names. Errors in segmentation were simulated by repeated applications of ‘erosion’ or ‘dilation’ operations to the true region of activation.

The degree of simulated segmentation error for each ROA was randomly selected from a standard normal distribution and rounded to the nearest integer. This random variable controlled the number of repetitions of the erosion/dilation operator – at least one repetition and at most two repetitions were applied. Thus the number of repetitions was an integer randomly chosen from the set  $[-2, -1, 1, 2]$  – the negative values indicated erosion and the positive values indicated dilation. For example, if the rounded random variable had the value -2, two successive applications of the erosion operator simulated the results of inaccurate segmentation of the ROA (for erosions, at least one pixel was retained in the ROA).

For comparison of classification accuracies with correct and incorrect segmentations, all ROAs in Test-Suite-Size were subjected to this erosion/dilation to simulate inaccurate segmentation. For the same image dataset, the classification accuracy from accurate segmentation was compared with that from inaccurate segmentation. The same features selection and machine learning procedures employed for the calculation of  $\text{Acc}_{\text{KDSf}}$  (Figure 27) was employed to calculate  $\text{Acc}_{\text{KDiSf}}$  – the classification accuracy for *incorrect* segmentation. Since features were constructed from incorrect boundaries of the regions of activation, the incorrect

sizes of the ROAs were used for machine learning. From the 12 repetitions for each generative model in Test-Suite-Size, the following hypothesis was tested:

$$\text{Acc}_{\text{KDSf}} > \text{Acc}_{\text{KDiSf}} \quad (\text{Hypothesis 4.3})$$

## 4.2.8 Machine Learning Details

### Neural Network

Feed-forward neural networks as implemented in Matlab [71] were used for this work. Two layer networks, with  $\sqrt{p}$  nodes in the single hidden layer (where  $p$  is the number of input features) were used for this purpose. So, for voxel-based machine learning (KBV),  $\sqrt{k}$  hidden nodes were used when  $k$ -best voxels were chosen as features (i.e. the hidden layer contained at most 64 nodes, when all the pixels were used in the model). Matlab implementation of the ‘resilient’ back-propagation algorithm (with 100 epochs) was used for training the networks.

### SVM

A publicly available Support Vector Machine implementation [43] with Matlab support was chosen for this work. Default options in this implementation (including ‘radial basis kernels’) were selected.

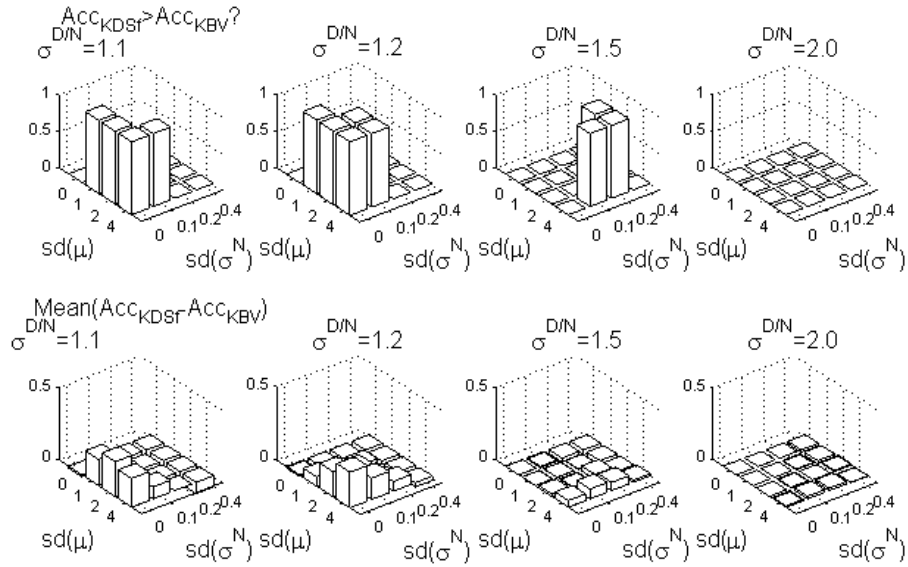
## 4.3 KDSF RESULTS

### 4.3.1 Test-Suite-Size (Without Smoothing)

#### 4.3.1.1 KDSf vs. KBV

For a narrow range of between-group differences in sizes of regions of activation (ROAs), the classification accuracy of KDSf was higher than that of KBV or PCA. However, this advantage of KDSf was *not* retained if the activation images were smoothed prior to feature construction for KBV and PCA. In this section, the accuracies of KDSf are compared with those for KBV and PCA without considering the effects of spatial smoothing – the impact of smoothing is considered in the next section.

In Test-Suite-Size, the  $\sigma^{D/N}$  parameter of the generative model (the ratio of mean ‘span’ of ROAs in disease group to that in normal group), was varied from 1.00 to 2.00. Hypothesis 4.1,  $\text{Mean}(\text{Acc}_{\text{KDSf}}) > \text{Mean}(\text{Acc}_{\text{KBV}})$ , was tested for different values of the parameters  $\text{sd}(\mu)$  and  $\text{sd}(\sigma^N)$ , controlling the location-variability and size-variability of the ROAs respectively. In all, 80 hypotheses were tested – one for each combination of the parameters of the generative model. The family-wise error rate for false discovery was controlled at  $p=0.05$  ( $p_{\text{adjusted}} = 0.05/80$ , using the Bonferroni correction for multiple comparisons). As shown in Figure 28, the hypothesis was accepted for  $\sigma^{D/N}$  values between 1.1 and 1.5 when the location-variability of the ROAs ( $\text{sd}(\mu)$ ) was non-zero.



**Figure 28.** Test-Suite-Size (no smoothing): Results for Hypothesis 4.1. The validity of the hypothesis  $\text{Mean}(\text{Acc}_{\text{KDSf}}) > \text{Mean}(\text{Acc}_{\text{KBV}})$ , for different conditions is shown in the top row. The bottom row shows the improvement in accuracy of KDSf over KBV.

With higher inter-subject variability of sizes of ROAs ( $\text{sd}(\sigma^N)$ ), the advantage of KDSf was muted – the increase in between-subject size-variability makes it more difficult for KDSf to detect the between-group difference in sizes. Somewhat surprisingly, the hypothesis was rejected in one case when inter-subject variability of size was 0 ( $\sigma^{D/N}=1.5$ ,  $\text{sd}(\mu)=4$  and  $\text{sd}(\sigma^N)=0$ ), however the p-value in this situation was 0.00093.

For the combinations of parameter values for which the hypothesis was accepted, the observed differences between the mean values of  $\text{Acc}_{\text{KDSf}}$  and  $\text{Acc}_{\text{KBV}}$  are shown in Table 7. As

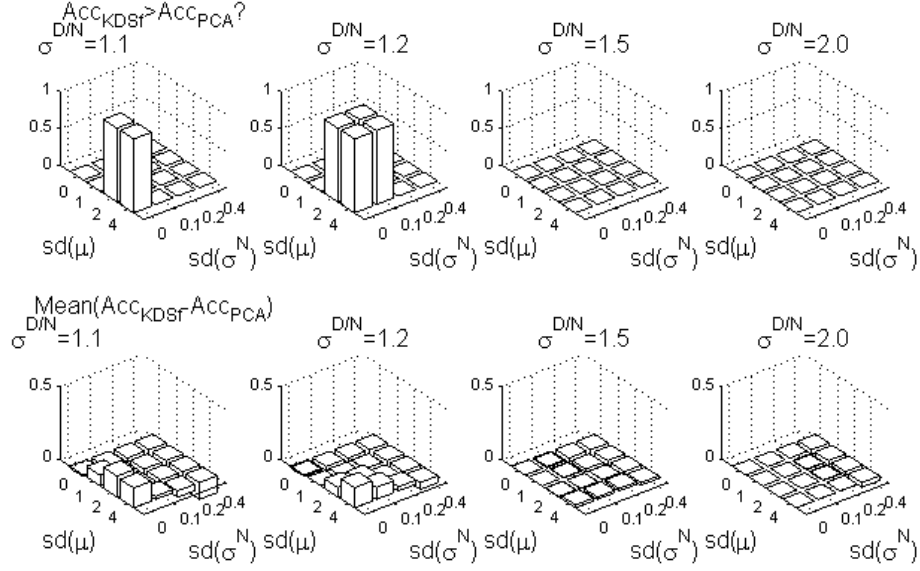
can be noted from the table, the accuracy improvements are more pronounced for smaller size differences between the two groups ( $\sigma^{D/N}=1.1$  and  $\sigma^{D/N}=1.2$ ).

**Table 7.** Test-Suite-Size (no smoothing): Improvements in classification accuracy of KDSf compared to that of KBV.

$\sigma^{D/N}$	$sd(\sigma^N)$	$sd(\mu)$	Mean(Acc <sub>KDSf</sub> )	Mean(Acc <sub>KBV</sub> )	Mean (Acc <sub>KDSf</sub> –Acc <sub>KBV</sub> )
1.1	0	1	0.911	0.732	0.179
1.1	0	2	0.904	0.667	0.238
1.1	0	4	0.857	0.660	0.197
1.1	0.1	4	0.697	0.615	0.082
1.2	0	1	0.978	0.917	0.061
1.2	0	2	0.979	0.808	0.171
1.2	0	4	0.935	0.693	0.242
1.2	0.1	2	0.896	0.769	0.126
1.2	0.1	4	0.842	0.697	0.144
1.5	0.1	4	0.956	0.889	0.067
1.5	0.2	2	0.953	0.917	0.036
1.5	0.2	4	0.904	0.849	0.056

#### 4.3.1.2 KDSf vs. PCA

For intermediate between-group differences in sizes of regions of activation (ROAs), the hypothesis regarding the improved classification accuracy of KDSf, compared to that of PCA, was found to be valid. The validity of Hypothesis 4.2,  $\text{Mean}(\text{Acc}_{\text{KDSf}}) > \text{Mean}(\text{Acc}_{\text{PCA}})$ , was tested for the same combination of generative model parameters as described in the previous section. As shown in Figure 29, for intermediate values of  $\sigma^{D/N}$  (1.1 and 1.2), the hypothesis was accepted. For these cases, the observed differences between the mean values of  $\text{Acc}_{\text{KDSf}}$  and  $\text{Acc}_{\text{PCA}}$  are shown in Table 8.



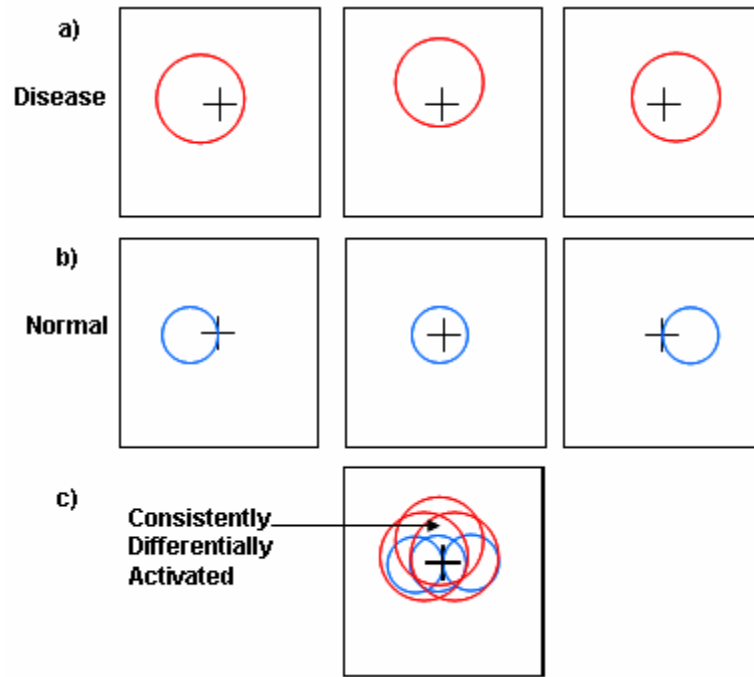
**Figure 29.** Test-Suite-Size (no smoothing): Results for Hypothesis 4.2. The validity of the hypothesis  $\text{Mean}(\text{Acc}_{\text{KDSf}}) > \text{Mean}(\text{Acc}_{\text{PCA}})$  for different conditions is shown in the top row. The bottom row shows the improvement in accuracy of KDSf over PCA.

**Table 8.** Test-Suite-Size (no smoothing): Improvement of classification accuracy of KDSf over PCA.

$\sigma^{D/N}$	$\text{sd}(\sigma^N)$	$\text{sd}(\mu)$	$\text{Mean}(\text{Acc}_{\text{KDSf}})$	$\text{Mean}(\text{Acc}_{\text{PCA}})$	$\text{Mean}(\text{Acc}_{\text{KDSf}} - \text{Acc}_{\text{PCA}})$
1.1	0	2	0.904	0.764	0.140
1.1	0	4	0.857	0.724	0.133
1.2	0	2	0.979	0.917	0.063
1.2	0	4	0.935	0.799	0.136
1.2	0.1	2	0.896	0.842	0.054
1.2	0.1	4	0.842	0.761	0.081

The results for Hypotheses 4.1 and 4.2 can be understood in terms of the effect of location-variability of the regions of activation on the reliability of individual voxels for discrimination between groups (see Figure 30). For intermediate between-group differences in sizes, the location-variability of the region of activation reduces the possibility that the set of voxels with different activation levels between the two groups is consistent between the training set and the testing set. However, for large differences in sizes, enough voxels are differentially activated to ensure adequate overlap between the training and testing sets, even in the presence of location-variability. Thus, the KBV approach is effective for large values of  $\sigma^{D/N}$  but not for intermediate values. Similarly, with PCA, the lack of sufficient overlap of the differentially

activated voxels leads to identification of components that are irrelevant for discrimination, leading to reduced accuracy (for intermediate values of  $\sigma^{D/N}$ ).



**Figure 30.** Differentially activated pixels. a) Examples of regions of activation for ‘disease’ images. b) Examples of regions of activation for ‘normal’ images c) Overlay of regions of activation shows the subset of pixels that are consistently activated in ‘disease’ but not in ‘normal’.

Grouping of voxels (either by PCA or KDSf) tends to reduce the dependence on individual differentially activated voxels, leading to improved accuracies. The KDSf approach provides better accuracy than PCA since the location-variability is directly accounted for. For high accuracy, the PCA approach requires that the differentially activated region in the image (see Figure 30) has similar locations in both the training set and the testing set.

While Hypotheses 4.1 and 4.2 are accepted for a narrow range of parameters of the generative model, the advantage of KDSf for these parameter-values is not retained if the activation images are smoothed prior to feature construction by PCA or KBV. The effect of smoothing on the hypothesis is described in the next section.

### 4.3.2 Test-Suite-Size (With Smoothing)

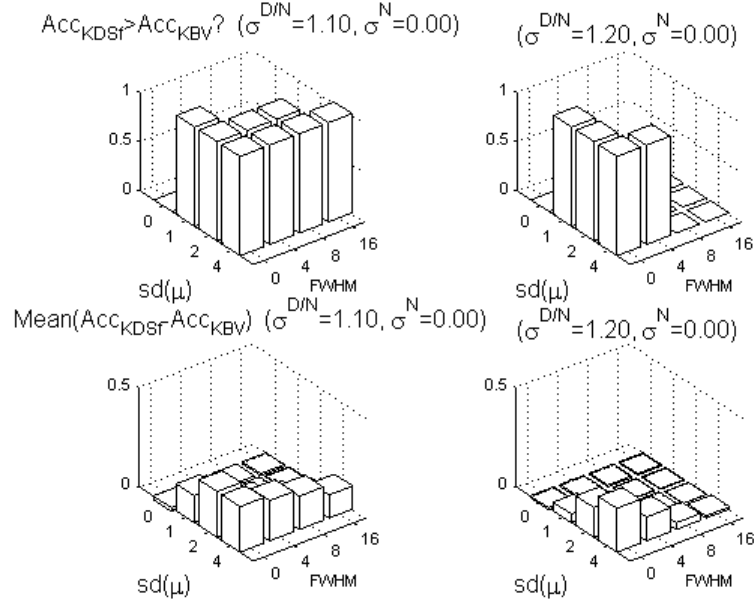
As shown in Figures 28 and 29, the accuracy of KDSf is consistently superior only when  $sd(\sigma^N) = 0$  – the accuracy suffers when  $sd(\sigma^N) = 0.1$  (see Tables 7 and 8 for  $\sigma^{D/N} < 1.5$ ). This subsection explores whether the advantage of KDSf at  $sd(\sigma^N)=0$  is retained if the images are smoothed prior to feature construction by KBV or PCA. Note that  $sd(\sigma^N)=0$  does not imply that there is no variation in sizes of ROAs between subjects, there is some variation due to the  $\rho_{xy}$  parameter in Table 5. The typical variability of sizes of ROAs for  $sd(\sigma^N)=0$  is listed in Table 9.

**Table 9.** Test-Suite-Size: Typical distribution of ROA sizes for  $sd(\sigma^N)=0$ .

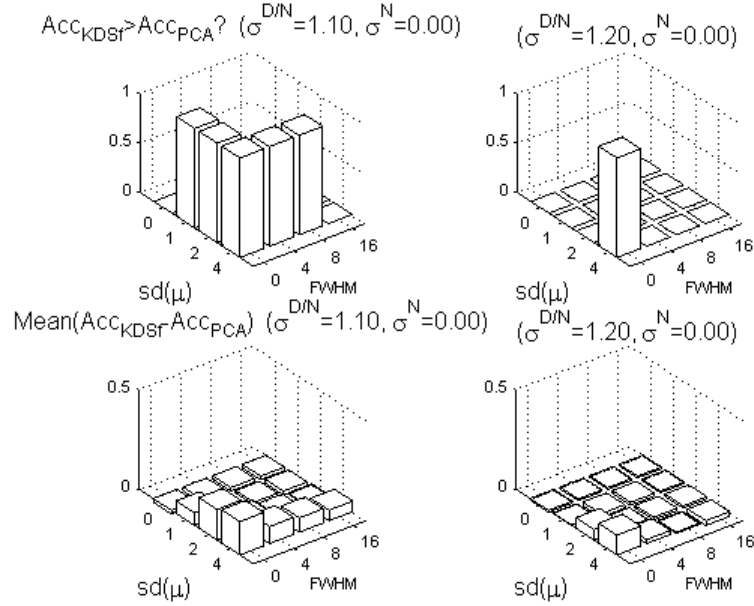
$\sigma^{D/N}$	Group	Max ROA size (pixels)	Mean ROA size (pixels)	Min ROA size (pixels)	Std. Dev. ROA size (pixels)
1.1	Normal	25	22.87	15	2.01
1.1	Disease	37	32.80	19	3.39
1.2	Normal	25	23.03	19	1.30
1.2	Disease	41	38.53	29	3.09

As can be observed from Figure 31, with the correct level of spatial smoothing, the KDSf accuracies are no longer significantly higher than those from KBV, except for extreme levels of location-variability ( $sd(\mu)=4$ ). Thus, except for extreme location-variability, Hypothesis 4.1 is rejected for some level of smoothing (FWHM).

Also, Figure 32 indicates that KDSf is no longer significantly more accurate than PCA, if the correct level of smoothing is introduced. For all levels of location-variability ( $sd(\mu)$ ), the hypothesis is rejected for some level of smoothing (FWHM). Table 10 shows that the accuracy of PCA with smoothed images is actually slightly higher than that of KDSf in some cases. This is because the between-subject size-variability makes it more difficult for KDSf to detect between-group differences in sizes – however, smoothing can blur some of the between-subject size-variability, allowing PCA to isolate the component characterizing the between-group differences. Note that, since smoothing causes loss of size information, KDSf does not use smoothed images for segmentation.



**Figure 31.** Test-Suite-Size: Effect of spatial smoothing on Hypothesis 4.1. The validity of the hypothesis  $\text{Mean}(\text{Acc}_{\text{KDSf}}) > \text{Mean}(\text{Acc}_{\text{KBV}})$ , for different conditions is shown in the top row. The bottom row shows the improvement in accuracy of KDSf over KBV.



**Figure 32.** Test-Suite-Size: Effect of spatial smoothing on Hypothesis 4.2. The validity of the hypothesis  $\text{Mean}(\text{Acc}_{\text{KDSf}}) > \text{Mean}(\text{Acc}_{\text{PCA}})$ , for different conditions is shown in the top row. The bottom row shows the improvement in accuracy of KDSf over PCA.



**Table 10.** Test-Suite-Size (with spatial smoothing): Accuracies for KDSf, KBV and PCA ( $\text{sd}(\sigma^N)=0$ ).

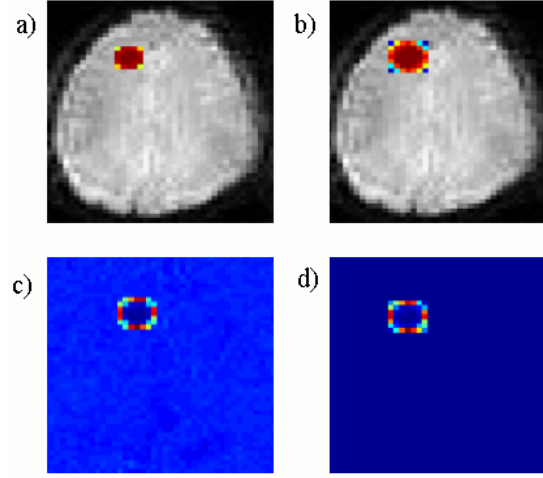
$\sigma^{D/N}$	$\text{sd}(\mu)$	FWHM	Mean $\text{Acc}_{\text{KDSf}}$	Mean $\text{Acc}_{\text{KBV}}$	Mean $\text{Acc}_{\text{PCA}}$	Mean ( $\text{Acc}_{\text{KDSf}}$ - $\text{Acc}_{\text{KBV}}$ )	Mean ( $\text{Acc}_{\text{KDSf}}$ - $\text{Acc}_{\text{PCA}}$ )
1.1	2	0	0.903	0.674	0.750	0.229	0.153
1.1	2	4	0.903	0.785	0.865	0.118	0.038
1.1	2	8	0.903	0.836	0.892	0.067	0.011
1.1	2	16	0.903	0.882	0.897	0.021	0.006
1.1	4	0	0.853	0.625	0.686	0.228	0.167
1.1	4	4	0.853	0.651	0.768	0.202	0.085
1.1	4	8	0.853	0.686	0.782	0.167	0.071
1.1	4	16	0.853	0.738	0.800	0.115	0.053
1.2	2	0	0.971	0.822	0.918	0.149	0.053
1.2	2	4	0.971	0.938	0.969	0.033	0.002
1.2	2	8	0.971	0.961	0.983	0.010	-0.012
1.2	2	16	0.971	0.972	0.982	-0.001	-0.011
1.2	4	0	0.915	0.696	0.817	0.219	0.098
1.2	4	4	0.915	0.796	0.896	0.119	0.019
1.2	4	8	0.915	0.882	0.921	0.033	-0.006
1.2	4	16	0.915	0.904	0.939	0.011	-0.024

While smoothing removes the advantage of KDSf for Test-Suite-Size, there are two difficulties associated with KBV and PCA for knowledge discovery. First, for smoothing to be effective for KBV and PCA, the optimal level of smoothing (FWHM) needs to be determined. Second, the features in KBV and PCA are clouds of voxels that need to be interpreted manually to infer that the sizes of ROAs are different between the two groups (see Figures 34 and 36, this may be even more difficult in 3D). Thus, given two feature selection methods of similar accuracy, the method with superior interpretability is preferred for knowledge discovery purposes.

#### 4.3.3 Test-Suite-Size: Effect of Location-variability

In the absence of location-variability of ROAs, the heat-map representation of the distribution of ROAs (Figures 33a and 33b) clearly shows the difference in sizes between the two groups. Each

pixel value in the heat-map image counts the number of subjects for which the pixel is activated – the variation within the ROA reflects the variability of shapes of ROAs in different subjects.



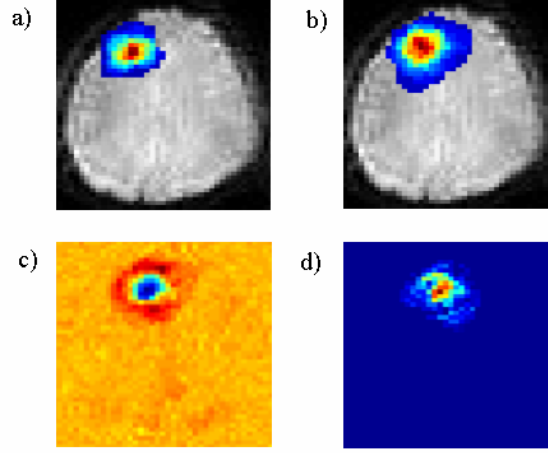
**Figure 33.** Test-Suite-Size: Feature selection in the absence of location-variability ( $\sigma^{D/N}=1.2$ ,  $sd(\sigma^N)=0$ ,  $sd(\mu)=0$ , and no smoothing). a) Heat-map image showing spatial distribution of ROA for the ‘normal’ group. b) Heat-map image for ‘disease’ group. c) Eigenimage 2 (as heat-map) identified by PCA. d) Interestingness values as heat-map: locations of the ‘best’ voxel features for KBV are shown in red.

PCA on this set of activation images yields a set of eigenimages (see section 2.4) – Figure 33c shows the eigenimage (as a heat-map) that is relevant for discrimination between the two groups. For KBV, the pixels are ranked based on an ‘interestingness’ measure (see section 2.4.1.2) – Figure 33d shows the interestingness measures for each pixel in the image (as a heat-map). In these heat-map images (Figures 33c and 33d), the ‘warmer’ colors represent higher values of component-membership and interestingness respectively.

In this case, both PCA-based feature selection (Figure 33c) and voxel-based feature selection (Figure 33d) correctly identifies the differentially activated voxels. Thus either method can successfully discriminate between the ‘normal’ and ‘disease’ images – however, absence of location-variability is unrealistic with current spatial normalization methods (see Chapter 2).

As shown in Figures 34a and 34b, with the introduction of location-variability (non-zero  $sd(\mu)$ ), the heat-map representation (showing the spatial distribution of the ROAs) is more dispersed. One of the eigenimages from PCA successfully identifies the pixels that reflect the size difference between the two groups (the cloud of red pixels around the ROA in Figure 34c). While PCA correctly identifies the relevant voxels in this case, this is not always the case (e.g. for  $\sigma^{D/N}=1.1$  this pattern of voxels was not identified as a component). The voxel-based (KBV)

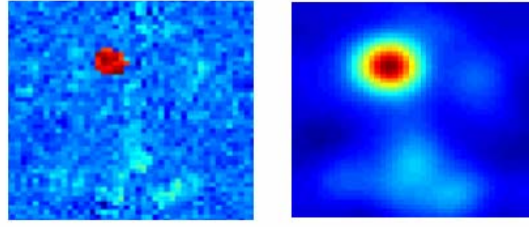
approach (Figure 34d) is unable to identify the voxels that reflect the size differences between the two groups. This explains why PCA is more competitive with KDSf than KBV (see Table 10).



**Figure 34.** Test-Suite-Size: Feature selection with location-variability ( $\sigma^{D/N}=1.2$ ,  $sd(\sigma^N)=0$ ,  $sd(\mu)=2$  and no smoothing). a) Heat-map image showing spatial distribution of ROA for the ‘normal’ group. b) Heat-map image for ‘disease’ group. c) Eigenimage 4 (as heat-map) identified by PCA. d) Interestingness values as heat-map: locations of the ‘best’ voxel features for KBV are shown in red.

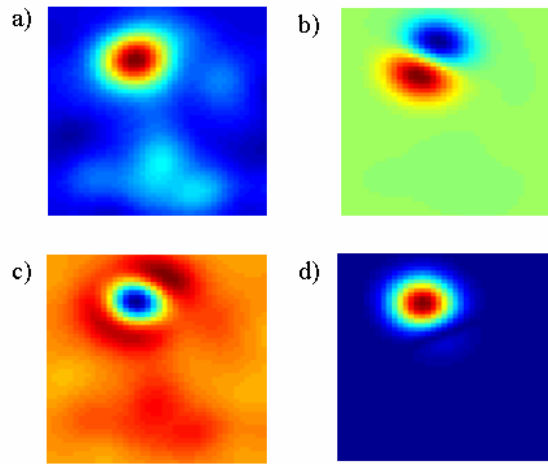
#### 4.3.4 Test-Suite-Size: Effects of Smoothing

Smoothing of the activation images prior to machine learning can reduce the dependence on small number of differentially activated voxels for discrimination between groups. Smoothing is achieved by convolution of a 2D Gaussian kernel with the activation image – the Full-Width-at-Half-Maximum (FWHM) parameter of the Gaussian kernel controls the degree of smoothing. Since smoothing ‘spreads’ the region of activation (see Figure 35), the effect of location-variability of ROAs is diluted – this increases the overlap between the differentially activated voxels in the training set and those from the testing set, leading to higher classification accuracies. The effect of smoothing on voxel-based feature selection and PCA-based feature selection is discussed next.



**Figure 35.** Smoothing an activation image (SPM). Heat-map representation of a simulated activation image before (left) and after smoothing (right) with a Gaussian kernel (FWHM=8). Higher values are shown with ‘warmer’ colors.

With spatial smoothing of the activation images followed by PCA, the discriminating eigenimage (Figure 36c) incorporates more discriminating voxels (the dark red voxels) than the eigenimage from unsmoothed images (Figure 34c). Machine learning with this (Figure 36c) eigenimage corresponds to consideration of evidence from a larger pool of discriminating voxels.

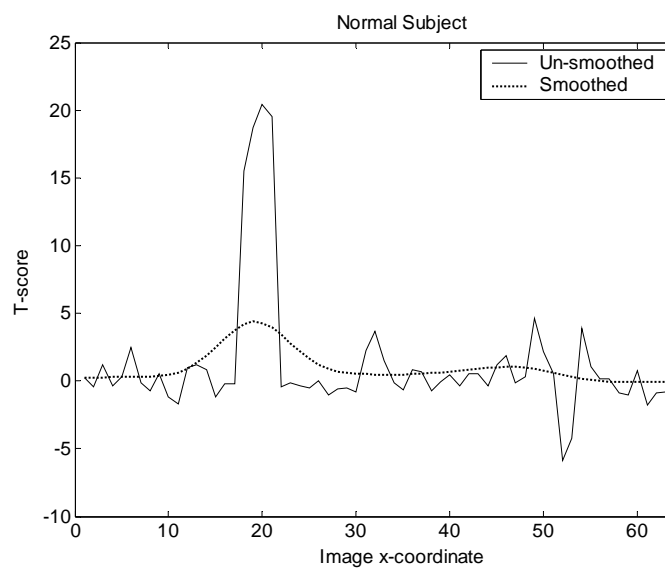


**Figure 36.** Test-Suite-Size: Effect of smoothing on feature construction in the presence of location-variability ( $\sigma^{D/N}=1.2$ ,  $sd(\sigma^N)=0$ ,  $sd(\mu)=2$  and FWHM=8). a) Eigenimage 1 from PCA. b) Eigenimage 2 from PCA. c) Eigenimage 4 from PCA. d) Locations of the ‘best’ voxel features for KBV (images smoothed prior to ‘interestingness’ calculations).

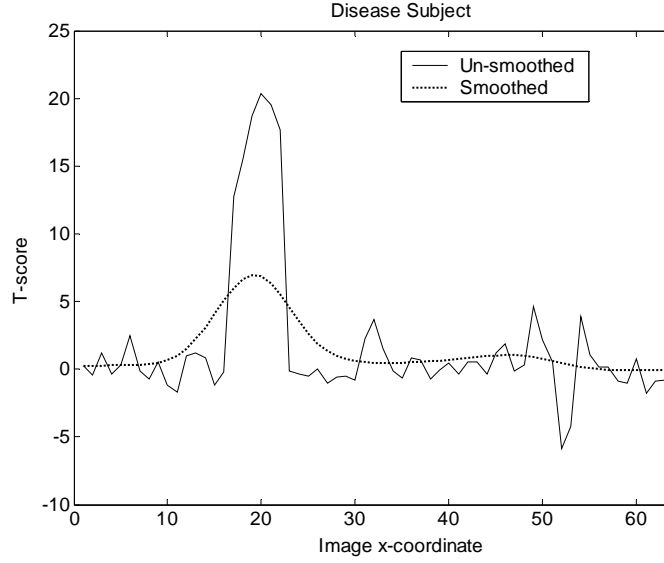
The effect of smoothing on voxel-based feature selection is two-fold. First, due to smoothing, more voxels in the neighborhood of the ROA are considered to be ‘interesting’ (compare Figures 34d and 36d). Second, the difference in ROA-sizes between the two groups introduces a difference in activation values (smoothed t-scores) between the two groups of smoothed images (compare smoothed t-scores between Figures 37 and 38). Note that prior to smoothing, the t-scores for the two groups were similar. This explains why the ‘interesting’ pixels (Figure 36d) are near the center of the ROA rather than at the periphery as for PCA

(Figure 36c). Since features (pixels) are ranked based upon mean differences in t-scores between the two groups (see section 2.4.1.2), the center voxels are more interesting because the center of the disease ROAs have higher smoothed t-scores due to the larger ROA-sizes in the disease group. Note that this is not the case in the absence of spatial smoothing and location-variability (Figure 33d) – in that case the peripheral pixels are correctly identified as ‘interesting’.

This distribution of the ‘interesting’ voxels near the center of the ROA also explains the higher classification accuracy of PCA compared to KBV. In KBV, even though the voxels at the center of the ROA are less relevant for discrimination than the voxels at the periphery of the ROA, the center voxels are considered first for feature selection.



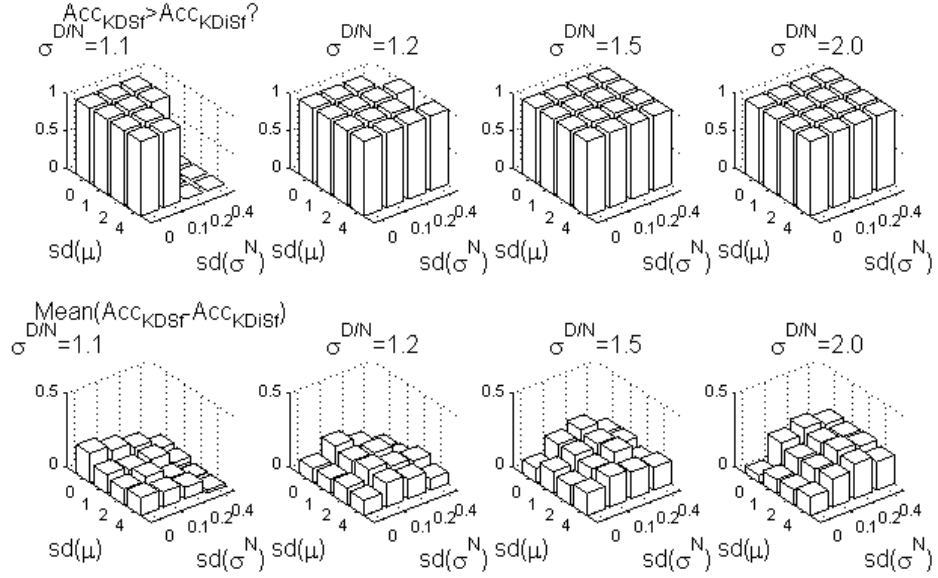
**Figure 37.** Test-Suite-Size: Effect of smoothing (FWHM=8) on activation image (SPM) for one ‘normal’ subject. The image is shown in profile for  $y=32$ . The peak is the region of activation.



**Figure 38.** Effect of smoothing (FWHM=8) on activation image (SPM) for one ‘disease’ subject. The image is shown in profile for  $y=32$ . The peak is the region of activation (note that ROA is wider for the ‘disease’ subject, compared to Figure 37).

#### 4.3.5 Test-Suite-Size: KDSf with inaccurate segmentation

The effect of segmentation-accuracy on classification-accuracy of KDSf was explicitly determined for all 80 situations in Test-Suite-Size. For this purpose, the KDSf accuracy with correct segmentation was compared with that from incorrect segmentation. The results for Hypothesis 4.3,  $\text{mean}(\text{Acc}_{\text{KDSf}}) > \text{mean}(\text{Acc}_{\text{KDiSf}})$ , are shown in Figure 39. The hypothesis was accepted for higher values of  $\sigma^{\text{D/N}}$ . For lower values of  $\sigma^{\text{D/N}}$ , the validity of the hypothesis depended upon the between-subject size-variability –  $\text{sd}(\sigma^{\text{N}})$ . This result can be understood in terms of the inherent difficulty of detecting small between-group differences in size, when the between-subject variability of size is also high. Thus, neither  $\text{Acc}_{\text{KDSf}}$  nor  $\text{Acc}_{\text{KDiSf}}$  achieved high accuracies for these cases.



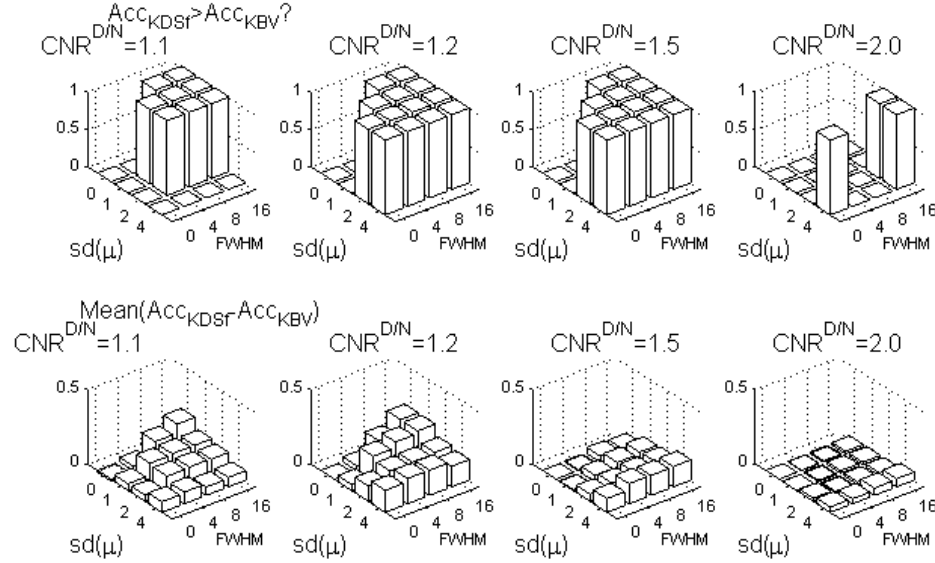
**Figure 39.** Test-Suite-Size: Results for Hypothesis 4.3. The validity of the hypothesis,  $\text{Mean}(\text{Acc}_{\text{KDSf}}) > \text{Mean}(\text{Acc}_{\text{KDiSf}})$ , for different conditions is shown in the top row. The bottom row shows the improvement in accuracy of KDSf over KDiSf.

### 4.3.6 Test-Suite-CNR

#### 4.3.6.1 KDSf vs. KBV

For intermediate ranges of between-group differences in activation levels (CNR), the hypothesis regarding the improved classification accuracy of KDSf, compared to that of KBV, was found to be valid. In Test-Suite-CNR, the  $\text{CNR}^{\text{D/N}}$  parameter of the generative model (the ratio of mean CNR in disease group to that in normal group), was varied from 1.00 to 2.00. Hypothesis 4.1,  $\text{Mean}(\text{Acc}_{\text{KDSf}}) > \text{Mean}(\text{Acc}_{\text{KBV}})$ , was tested for different values of the parameters  $\text{sd}(\mu)$  and FWHM – the former modeling the location-variability of the ROAs and the latter compensating for it by spatial smoothing. In all, 80 hypotheses were tested – one for each combination of the parameters. The family-wise error rate for false discovery was controlled at  $\dot{\alpha}=0.05$  ( $\dot{\alpha}_{\text{adjusted}} = 0.05/80$ , using the Bonferroni correction for multiple comparisons). As shown in Figure 40, the hypothesis was accepted for  $\text{CNR}^{\text{D/N}}$  values between 1.2 and 1.5, when  $\text{sd}(\mu)$  was greater than 1. For combinations of parameter values for which Hypothesis 4.1 was accepted, regardless of the degree of smoothing, the mean values of  $\text{Acc}_{\text{KDSf}}$  and  $\text{Acc}_{\text{KBV}}$  are

shown in Table 11. For lower levels of between-group differences ( $\text{CNR}^{D/N}=1.1$ ), neither KDSf nor KBV were very accurate – however smoothing masked true differences in the activation levels, leading to lower accuracies for KBV.



**Figure 40.** Test-Suite-CNR: Results for Hypothesis 4.1 (‘size’ variability,  $\text{sd}(\sigma^N)=0.1$ ). The validity of the hypothesis  $\text{Mean}(\text{Acc}_{\text{KDSf}}) > \text{Mean}(\text{Acc}_{\text{KBV}})$ , for different conditions is shown in the top row. The bottom row shows the improvement in accuracy of KDSf over KBV.

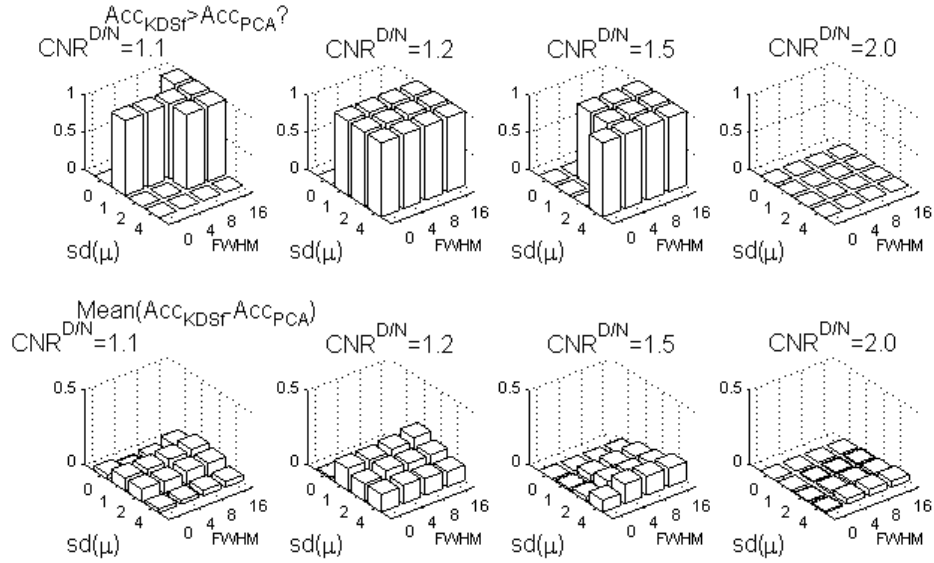
With higher location-variability of the ROAs (higher  $\text{sd}(\mu)$ ), smoothing was not sufficient for KBV to achieve the same accuracy as KDSf. This is because even as smoothing increases the pixel overlap between ROAs (and compensates for location-variability), smoothing also reduces the activation strength at the expanded periphery of the ROAs (see Figures 35 and 37) – this can mask inter-group differences in activation levels. This undesirable effect of smoothing on KBV can be seen for  $\text{CNR}^{D/N}$  values from 1.1 to 1.5 – excessive smoothing can cause  $\text{Acc}_{\text{KBV}}$  to suffer even when location-variability is absent ( $\text{sd}(\mu)=0$ ).

Thus, for between-group differences in activation level ranging from 20% to 50%, in the presence of location-variability of ROAs, KDSf can provide better classification accuracies than KBV (with or without spatial smoothing). This shows that the conventional dependence on spatial smoothing to account for lack of precise spatial normalization leads to loss of classification accuracy – spatial smoothing may mask substantial inter-group differences in activation levels.



#### 4.3.6.2 KDSf vs. PCA

As shown in Figure 41, for intermediate ranges of between-group differences in activation levels (CNR), the hypothesis regarding the improved classification accuracy of KDSf, compared to that of PCA, was found to be valid. The validity of Hypothesis 4.2,  $\text{Mean}(\text{Acc}_{\text{KDSf}}) > \text{Mean}(\text{Acc}_{\text{PCA}})$ , was tested for the same combination of generative model parameters as described in the previous section. Again, for intermediate values of  $\text{CNR}^{\text{D/N}}$  (for 1.2 and, to a lesser degree, for 1.5), the hypothesis was found to be valid. For these values of  $\text{CNR}^{\text{D/N}}$  (at higher levels of location-variability) the mean values of  $\text{Acc}_{\text{KDSf}}$  and  $\text{Acc}_{\text{PCA}}$  are compared in Table 11.



**Figure 41.** Test-Suite-CNR: Results for Hypothesis 4.2 ('size' variability,  $\text{sd}(\sigma^{\text{N}})=0.1$ ). The validity of the hypothesis  $\text{Mean}(\text{Acc}_{\text{KDSf}}) > \text{Mean}(\text{Acc}_{\text{PCA}})$  for different conditions is shown in the top row. The bottom row shows the improvement in accuracy of KDSf over PCA.

**Table 11.** Test-Suite-CNR: Classification accuracy of KDSf compared to that of KBV and PCA.

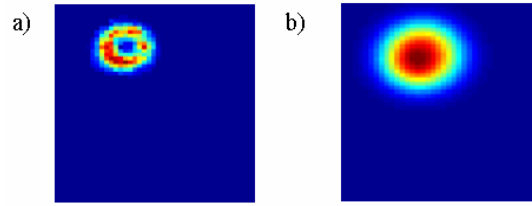
<b>CNR<sup>D/N</sup></b>	<b>sd(<math>\mu</math>)</b>	<b>FWHM</b>	<b>Mean Acc<sub>KDSf</sub></b>	<b>Mean Acc<sub>KBV</sub></b>	<b>Mean Acc<sub>PCA</sub></b>	<b>Mean (Acc<sub>KDSf</sub> - Acc<sub>KBV</sub>)</b>	<b>Mean (Acc<sub>KDSf</sub> - Acc<sub>PCA</sub>)</b>
1.20	2	0	0.890	0.789	0.774	0.101	0.117
1.20	2	4	0.890	0.746	0.742	0.144	0.149
1.20	2	8	0.890	0.742	0.750	0.149	0.140
1.20	2	16	0.890	0.735	0.750	0.156	0.140
1.20	4	0	0.836	0.689	0.717	0.147	0.119
1.20	4	4	0.836	0.667	0.713	0.169	0.124
1.20	4	8	0.836	0.668	0.724	0.168	0.113
1.20	4	16	0.836	0.699	0.735	0.138	0.101
1.50	2	0	0.997	0.949	0.968	0.049	0.029
1.50	2	4	0.997	0.933	0.921	0.064	0.076
1.50	2	8	0.997	0.917	0.924	0.081	0.074
1.50	2	16	0.997	0.907	0.919	0.090	0.078
1.50	4	0	0.990	0.899	0.911	0.092	0.079
1.50	4	4	0.990	0.856	0.860	0.135	0.131
1.50	4	8	0.990	0.857	0.863	0.133	0.128
1.50	4	16	0.990	0.854	0.867	0.136	0.124

#### 4.3.7 Conclusions from KDSf simulation study

The goal of this simulation study was to determine if the KDSf approach is worthwhile – that is, if there are some types of between-group differences in regions of activations (ROAs) for which the conventional methods of feature construction may be inadequate for automated knowledge discovery. From Test-Suite-Size, it is concluded that while KDSf can improve accuracy for a narrow range of differences in ROA-sizes, this advantage is lost if the activation images are smoothed prior to feature selection with PCA and KBV. Thus, hypothesis 4.1 and 4.2 are rejected for this situation when the two groups differ with respect to sizes of ROAs. However, it should be noted that KDSf accuracy remains competitive with other methods for all cases tested here.

However, from Test-Suite-CNR, it is determined that for a range of CNR differences between the two groups, KDSf can provide higher classification accuracies than KBV and PCA, regardless of smoothing. Thus hypothesis 4.1 and 4.2 are accepted for this case.

In summary, it has been demonstrated that, in some circumstances, the KDSf approach to feature construction can be helpful for knowledge discovery from fMRI datasets. It should be noted however that the classification accuracies reported here do not reflect the prevalence of the disease condition in the general population – thus classification accuracies reported during knowledge discovery may not be reproduced in clinical practice. It is also demonstrated that voxel-based feature selection with prior smoothing of images (KBV) can be misleading regarding the nature of difference between the two groups – differences in sizes of ROA may be misinterpreted as differences in activation strengths (see Figure 42). This finding is also applicable to conventional voxel-based hypothesis-testing approaches to knowledge discovery (high t-scores at the center of the consensus ROA in the group SPM can be misleading).



**Figure 42.** Test-Suite-Size: Effect of smoothing on voxel-based feature construction in the presence of location-variability ( $\sigma^{D/N}=2$ ,  $sd(\sigma^N)=0.1$ ,  $sd(\mu)=1$ ). a) Without smoothing (FWHM=0), locations of ‘interesting’ voxels correctly suggest a ‘size’ difference between groups. b) With smoothing (FWHM=8), locations of ‘interesting’ voxels incorrectly suggest that the groups differ in strengths of activation.

While smoothing removes the accuracy advantage of KDSf in some cases, knowledge discovery with KDSf directly identifies the attribute of ROAs that is different between the two groups (e.g. size differences or CNR differences). The clouds of discriminating voxels identified with PCA and KBV may be difficult to interpret for knowledge discovery purposes (particularly in 3D). Thus, the superior interpretability of KDSf features, and high classification accuracies obtained with KDSf are some of the advantages of the KDSf framework for automated knowledge discovery from fMRI datasets.

## 5.0 SEGMENTATION WITH ACEIC

The KDSf framework requires an automated method for functional segmentation of fMRI images which can identify the boundaries of the regions of task-induced activation. In this Chapter functional segmentation of an fMRI image is defined as the partitioning of the image voxels into clumps that are co-modulated by task-related neural activation or other processes (a discussion of other possibilities for functional segmentation is presented in Chapter 4). In this Chapter, the goal is to correctly isolate a set of contiguous voxels that share similar activation timecourses. In functional clustering [72], the goal is to group similar time-courses from any region of the brain – the voxels in a group need not be contiguous. Functional segmentation differs from functional clustering in the additional requirement of spatial connectivity of the voxels in a partition (group). Note however that connected sub-clusters from clustering solutions can be interpreted as segments. The accuracy of extant clustering methods is dependent on the appropriateness of the user-specified clustering parameters – such as number of clusters or cluster homogeneity thresholds [51]. This work is motivated by the desire for an accurate method that does not require the user to specify such parameters, which can be difficult to choose for arbitrary images.

In this chapter, a new method for functional segmentation is presented – Auto-threshold Contrast Enhancing Iterative Clustering (or ACEIC) that is based upon maximization of the contrast (differentiation) of an image-region with respect to its spatial neighborhood (region-contrast). Segmentation with maximization of region-contrast is appealing since no prior information about appropriate number of clusters or cluster homogeneity thresholds is required. While contrast-maximizing methods have been used for segmentation of intensity images [64], contrast-maximization has not been applied to fMRI images. During the course of this work, it was observed that contrast measures used with intensity images do not work well for timecourse-valued images (there is no clear peak for the correct segmentation solution). For this reason, a set

of contrast measures were empirically evaluated and a contrast measure that was effective for fMRI images was chosen from this set.

Region-growing techniques, such as greedy agglomeration, have been used along with contrast maximization for segmentation of intensity images [64]. However, with timecourse-valued images with low levels of neural activation, greedy agglomeration with maximization of region-contrast can lead to under-segmentation – the global maximum of region-contrast is typically not at the correct segmentation solution (see below). To address this problem, in this work, the region-growing process is generalized from greedy agglomeration of one-voxel at-a-time to agglomeration controlled by a region homogeneity threshold. The region homogeneity threshold controls the variety of timecourses that can be included inside a region – thus indirectly controlling the size of the region. Instead of finding the region-size with the best contrast, the segmentation problem is framed in terms of finding the homogeneity threshold beyond which there is no appreciable contrast.

In this approach to region-growing, the current region is grown by clustering the voxels currently inside the region with the voxels neighboring the region – each cluster is constrained to satisfy the specified region homogeneity threshold. Each cluster encountered during region-growing is a potential segmentation solution – of these, the solution with the highest region-contrast is retained. This iterative homogeneity-constrained region-growing process is called Contrast Enhancing Iterative Clustering or CEIC. With the proper homogeneity threshold, this approach can solve the under-segmentation problem associated with greedy agglomeration. However, the proper homogeneity threshold still needs to be identified.

This is done automatically, using a method (ACEIC) that searches for the optimal homogeneity threshold to be used by CEIC. In this approach, the segmentation problem is defined in terms of finding the homogeneity threshold beyond which region-contrast deteriorates. The ACEIC method employs an iterative grid-search to identify this optimal homogeneity threshold. The main user-specified parameter required by the ACEIC method is an upper limit on region size, which can be chosen based upon available computational capacity and biological knowledge about sizes of activated regions in the brain.

In the next section, some of the prior work related to functional segmentation – including Probabilistic ICA – is described. Probabilistic ICA (PICA) is a popular exploratory method that also does not require user-specified parameters such as number of components or significance

thresholds. In later sub-sections, the CEIC and ACEIC methods are described in detail. Then, the accuracy of the ACEIC method is compared with that of PICA for an independently published benchmark dataset. Finally, the ACEIC method is applied to some data from an actual fMRI experiment. It is observed that while ACEIC can be used for functional segmentation, there may be some fracturing of functional units for lower levels of neural activation. It is also demonstrated that segmentation with ACEIC can help to identify signal artifacts which can confound regression-based analysis.

## 5.1 RELATED WORK

As described in Chapter 2, three main approaches have been proposed for grouping voxels based on similarity of timecourses – clustering [48-51], Independent Component Analysis [52, 53, 61] and image segmentation methods [55]. Of these, the first two approaches do not restrict the voxels within a cluster (or component) to be spatially connected.

One problem with clustering methods is the need for user-specified parameters (e.g. initial number of clusters or level of the cut in the hierarchy) which determine the quality of the clustering results. Recently, a comparative analysis of various clustering methods has shown that the accuracy of clustering results is typically dependent upon the appropriate choice of parameters and upon random initializations [51].

The goal of Independent Component Analysis (ICA) [52] of fMRI data is to find a set of independently distributed spatial patterns which, when linearly combined with a square mixing matrix, reproduces the observed data. Each component is characterized by a timecourse pattern, one of which typically corresponds to the fMRI task. Each spatial component specifies a weight for each voxel denoting the degree of participation of the voxel in the component. However, these weights must be thresholded to label a subset of the voxels as ‘significant’ participants in the component – ICA requires the user to specify this threshold.

Probabilistic ICA [53] allows for non-square mixing with Bayesian estimation of the model order (number of components). In addition, in Probabilistic ICA, the independent component maps are assessed for significance with a Gaussian mixture model, without requiring

the user to specify a threshold. Note that the accuracy of ICA (or PICA) is dependent upon the validity of the model assumptions [61].

Image processing approaches such as seeded region-growing [55] also require the user to specify a region homogeneity threshold. The region-growing is constrained by the requirement that a homogeneity measure computed for the region satisfies a user-defined threshold. However, given the variety of noise sources in fMRI data, it is difficult to choose the optimal homogeneity threshold for individual images. Other typical image segmentation approaches, such as split-and-merge segmentation [63], are also dependent upon the proper choice of region homogeneity thresholds.

Greedy region-growing with maximization of region-contrast has been used for intensity-valued medical images [64]. While contrast-maximization does not require user-specified parameters, as mentioned earlier, contrast measures used with intensity-valued images may not be suitable for timecourse-valued images. Further, greedy region-growing can lead to under-segmentation for image regions with lower levels of neural activation.

This work is motivated by the desire for an accurate segmentation method that does not require dataset-specific parameters such as number of clusters or homogeneity thresholds. Maximization of region-contrast along with a generalization of greedy region-growing is employed in this work to achieve these goals.

## 5.2 DETAILS OF ACEIC

Segmentation of an fMRI dataset by maximization of region-contrast requires the selection of a distance (or dissimilarity) metric to compare timecourses for voxels. Also, based upon the distance metric, a region-contrast measure needs to be chosen.

In this work, the distance measure chosen to compare timecourses is based upon correlation, which is commonly used to assess the similarity of timecourses [73]. The correlation distance  $d_{vw}$  between the signal timecourses for two voxels  $v$  and  $w$  is defined based upon the correlation between the row vectors  $\vec{x}_v$  and  $\vec{x}_w$  representing the signal values for voxels  $v$  and  $w$  for the  $T$  time-points.

$$d_{vw} = 1 - r_{vw} \quad (5.1)$$

where  $r_{vw}$  is the Pearson correlation coefficient between the timecourses  $\bar{x}_v$  and  $\bar{x}_w$ .

While a variety of contrast measures have been used for intensity images, many of them do not work well for timecourse-valued images. For this reason the characteristics of a few contrast measures were empirically studied and a contrast measure that incorporated peripheral contrast and region heterogeneity was chosen for use with fMRI images – these terms are defined below.

A region (or segment)  $R$  is a non-empty set of spatially connected voxels using the 26-connected neighborhood model for 3D images. The spatial connectivity of  $R$  requires that for any two voxels in the set  $R$ , there must be a path between the two voxels such that all the voxels in the path belong to  $R$  and all sequential pairs of voxels in the path are 26-connected. In other words, a region cannot consist of islands of voxels, but a region can be hollow. If  $v$  and  $w$  are two voxels, the neighbor relationship is defined as  $N(v, w) = 1$  iff  $v$  and  $w$  are 26-connected. The interior boundary  $I$  is the subset of voxels in  $R$  that have at least one neighbor voxel which is not in the set  $R$ .

$$I = \{v \mid v \in R, w \notin R, N(v, w)\} \quad (5.2)$$

The exterior boundary  $E$  is the set of voxels outside the set  $R$  that have at least one neighbor voxel in the set  $R$ .

$$E = \{v \mid v \notin R, w \in R, N(v, w)\} \quad (5.3)$$

Peripheral Contrast  $c_P$  is the average dissimilarity between neighbor voxels from the interior boundary and the exterior boundary.

$$c_P = \text{mean}\{d_{vw} \mid v \in I, w \in E, N(v, w)\} \quad (5.4)$$

Region Heterogeneity  $h_R$  is a measure of the average dissimilarity of the timecourses within a region  $R$ .

$$h_R = \text{mean}\{d_{vw} \mid v \in R, w \in R, v \neq w\} \quad (5.5)$$

Exterior Boundary Heterogeneity  $h_E$  is a measure of the average dissimilarity of the timecourses within the Exterior Boundary  $E$ .

$$h_E = \text{mean}\{d_{vw} \mid v \in E, w \in E, v \neq w\} \quad (5.6)$$



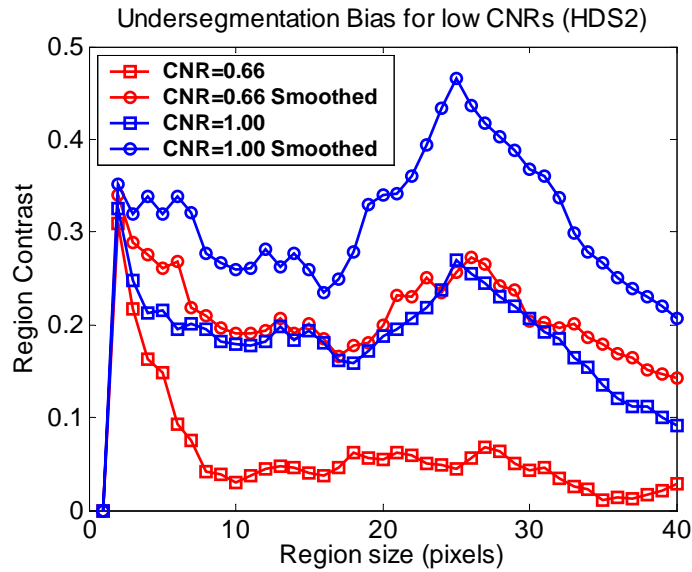
The Region ‘Width’  $w_R$  is the maximum dissimilarity of timecourses within the region  $R$ .

$$w_R = \max\{d_{vw} \mid v \in R, w \in R, v \neq w\} \quad (5.7)$$

The Region Homogeneity Criterion (RHC) is the requirement that  $w_R$  for a Region  $R$  is not greater than a specified threshold. This threshold is called the Region Homogeneity Threshold (RHT).

Finally, the region-contrast measure  $c_R$  is defined as the Peripheral Contrast  $c_P$  offset by the Region Heterogeneity  $h_R$ . However, since  $h_R$  is not defined for singleton regions, in this case,  $h_R$  is approximated by the Exterior Boundary Heterogeneity  $h_E$ .

$$c_R = \begin{cases} \max\{c_P - h_R, 0\} & \text{if } |R| > 1 \\ \max\{c_P - h_E, 0\} & \text{if } |R| = 1 \end{cases} \quad (5.8)$$



**Figure 43.** Under-segmentation problem associated with contrast maximization with greedy agglomeration.

The global maximum may be away from the correct solution (at 25 pixels).

Figure 43 shows typical plots of this region-contrast measure ( $c_R$ ) as a function of the steps in the greedy region-growing process. The contrast-to-noise ratios (CNR values) shown in the figure reflect marginally detectable activation-induced increase in the signal intensity, relative to the characteristic noise level of the timecourses. As shown in the figure, for regions with lower activations, maximization of region-contrast with greedy agglomeration can lead to

under-segmentation – temporal smoothing [74] of the timecourses provides only partial relief. The CEIC method described in the next subsection is designed to overcome this problem.

### 5.2.1 CEIC

As mentioned earlier, for lower activation levels, maximization of region-contrast with greedy agglomeration can lead to under-segmentation. As a generalization of greedy voxel-by-voxel agglomeration, the Contrast Enhancing Iterative Clustering (CEIC) method (Figure 44) is developed – this approach grows regions by homogeneity-constrained iterative clustering.

Starting with a seed voxel, iterative clustering is used to sample the space of possible region definitions, constrained by the particular region homogeneity threshold. For larger homogeneity thresholds, more voxels are added at a time, effectively stepping over the high-contrast region-definitions associated with under-segmentation. In each iteration, the union of the voxels in the current region-definition and its immediate neighbors (Exterior Boundary  $E$ ) are clustered (hierarchical clustering with complete linkage) such that each cluster satisfies the homogeneity threshold. The cluster containing the seed voxel is pruned to remove any member voxels that are not spatially connected with the seed voxel. This pruned cluster acts as the new candidate region-definition which is then used in the next iteration. Thus, starting with a seed voxel, the CEIC method iteratively samples the space of possible region-definitions that satisfy the homogeneity threshold until the region-definition converges.

**Algorithm:** CEIC

**Configuration parameters:** distance metric, contrast measure, maximum region size

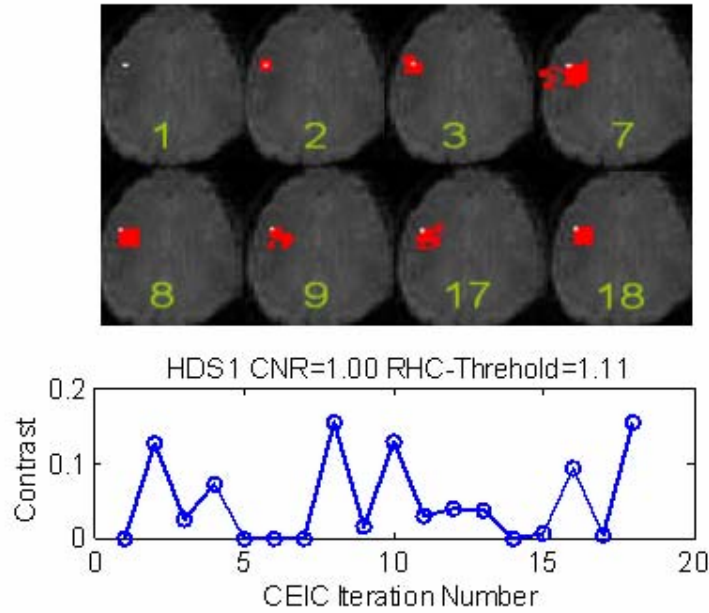
**Inputs:** seed voxel, RHT (region homogeneity threshold)

**Outputs:** final region-definition

*begin*

0. set *region\_definition* to singleton region with the seed voxel
1. set *region\_contrast* to the computed contrast for *region\_definition*
2. cache the *region\_definition* and the corresponding *region\_contrast*
3. set *done* to false
4. while not *done*
  5. set *neighbors* to the set of still un-claimed voxels in exterior boundary  
of *region\_definition*
  6. maintaining RHT, cluster (hierarchical with complete linkage) the union  
of *region\_definition* and *neighbors*
  7. set *new\_region\_definition* to the spatially connected sub-cluster containing  
the seed-voxel
  8. if *new\_region\_definition* is already present in the cache
  9. set *termination\_condition* to *CONVERGENCE*, and set *done* to true
  - end
  10. if size of *new\_region\_definition* exceeds the maximum region size parameter
  11. or if the region growth shows oscillations (this check is optional)
  12. set *termination\_condition* to *SIZE*, and set *done* to true
  - end
  13. set *region\_definition* to *new\_region\_definition*
  14. set *region\_contrast* to the computed contrast for *region\_definition*
  15. cache the *region\_definition* and the corresponding *region\_contrast*
  - end
  16. return the region in the cache with maximum region-contrast as the final region-definition
- end*

**Figure 44.** Algorithm CEIC – Contrast Enhancing Iterative Clustering.



**Figure 45.** Convergence of CEIC (iterations 8 and 18 are the same). The contrasts for the candidate region-definitions are also shown.

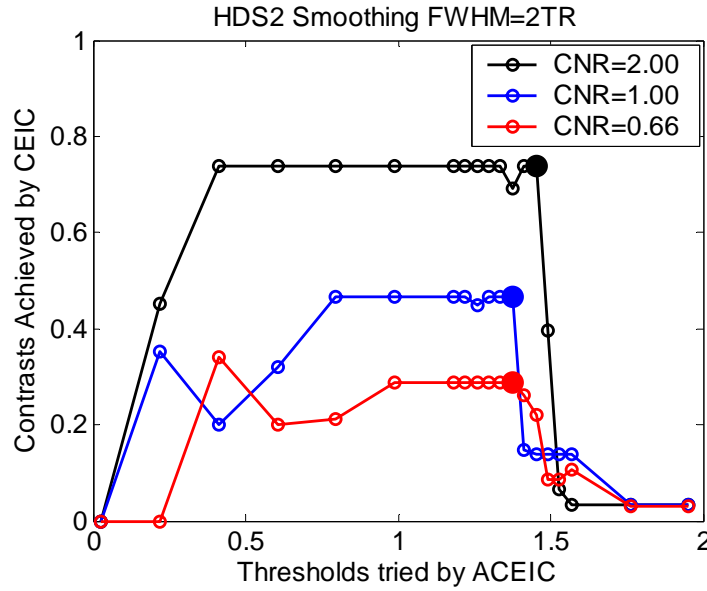
Convergence (Figure 45) is reached when a previously encountered region-definition is seen again – further iterations will not generate new region-definitions. Since the search space of possible region-definitions is finite, convergence is guaranteed, but it can be time-consuming for large search-spaces such as 3D regions with low levels of activation. Note that, unlike the greedy-agglomeration method, the CEIC method is non-monotonic – iteration 3 in Figure 45 includes a pre-existing structure near the true region of activation, but these incorrect pixels are discarded as more homogeneous pixels become available in later iterations.

For practical considerations, stopping criteria other than convergence are necessary. A user-specified upper limit on region size is utilized as an additional stopping criterion. The choice of this parameter can be based upon available computational capacity and biological knowledge about sizes of activated regions. Additionally, for 3D images where the search space is large, the region-definition may oscillate for regions with weak activation. In this case, if the region-size stagnates or falls significantly, the search can be terminated.

For a given region homogeneity threshold, when the stopping criterion is reached, the CEIC method chooses the maximum contrast (maximum  $c_R$ ) solution from the candidate region-definitions encountered during the search. For the final segmentation solution, the next step is the identification of the optimal region homogeneity threshold for the given seed voxel.

### 5.2.2 ACEIC

To automate the selection of the optimal region homogeneity threshold, the Auto-threshold CEIC (ACEIC) method is developed. The goal of the ACEIC method is to determine the homogeneity threshold beyond which there is no appreciable region-contrast. Figure 46 shows the contrast achieved by the CEIC method at different values of the homogeneity threshold – the results for three synthetic images with different activation levels are shown. Note that for contrast-to-noise-ratio (CNR) of 0.66 (weak activation), choosing the threshold associated with the global maximum of region-contrast will lead to under-segmentation. For each CNR, the optimal threshold is at the right edge of the plateau of contrast values in the figure. To identify the right edge of the plateau, ACEIC employs an iterative grid-search of the possible range of thresholds to find the threshold that minimizes the forward-difference of the achieved contrast.



**Figure 46.** Automated threshold selection by the ACEIC method. The optimal thresholds are shown with filled symbols.

If  $c_R(t)$  is the best contrast achieved by the CEIC method for a homogeneity threshold  $t$ , the threshold  $t'$  that minimizes the forward difference operator for  $c_R(t)$  is identified – the loss of contrast is maximal at this threshold. An iterative grid-search over the range  $(0 < t \leq \tau)$  of possible homogeneity thresholds is used to identify  $t'$  – here  $\tau$  is the maximum possible value for the distance metric  $d_{vw}$ .

$$t' = \arg \min_{0 < t \leq \tau} (c_R(t + \Delta t) - c_R(t)) \quad (5.9)$$

The optimal threshold, beyond which contrast deteriorates, is at the right edge of the plateau of contrast values. However, for lower CNRs, this optimal threshold may not be the same as  $t'$  (Figure 46). The optimal threshold is approximated by the threshold  $t''$  ‘near’  $t'$  that has maximum contrast (shown with filled symbols in Figure 46).

$$t'' = \arg \max_{t' - \tau/10 < t \leq t'} (c_R(t)) \quad (5.10)$$

This final threshold  $t''$  identified by the ACEIC method is used to determine the final segmentation solution for the specified seed voxel.

Since ACEIC is a seeded region-growing method, multiple invocations of the ACEIC method are needed to segment the full image – the order of seed selection is an important consideration. While the ACEIC method is generally robust for alternate choices of seeds (see section 5.3.1.2), the segmentation results are somewhat dependent upon the order of seed selection, particularly for regions with weak activation (lower CNR). By default, the voxels are sorted by the maximum height of spectral peaks of the timecourses and are used as seeds in this order. Once a set of voxels is claimed by a segment, these voxels become invisible to the continuing segmentation process and do not participate in future neighbor or contrast calculations.

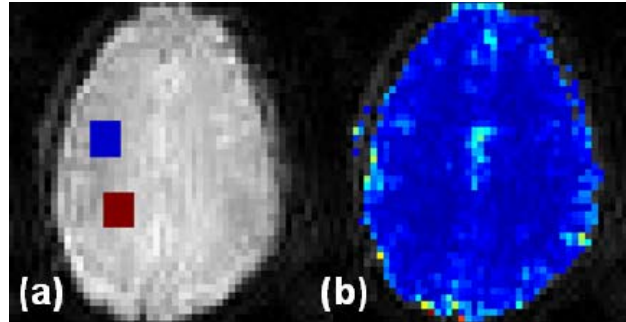
## 5.3 EVALUATION OF ACEIC

### 5.3.1 Evaluation with Benchmark Data

#### 5.3.1.1 Datasets HDS1/HDS2

An independently published benchmark fMRI dataset [51] has been used to compare various clustering techniques such as k-means, hierarchical clustering and fuzzy clustering. This hybrid (synthesized) dataset superimposes artificial activation patterns to a baseline in vivo MRI dataset. The baseline dataset consisted of time-series of 140 images with a matrix size of 64x64 pixels. To simulate task-related activation, the timecourses for the pixels in a 5x5 square region

of the slice were modulated by a box-car activation pattern (20 on, 20 off, repeated 3 times). As noted earlier, this hybrid approach is more realistic than datasets created from mathematical models of MR signal characteristics (mathematical phantoms).



**Figure 47.** a) Locations of artificial regions of activation in HDS1 (upper square) and HDS2 (lower square). b) Locations of pre-existing structures in the baseline image are highlighted in the heat-map image, where ‘warmer’ colors indicate the presence of strong periodic components in the timecourses.

The original dataset featured three different activation levels corresponding to contrast-to-noise ratios (CNR) of 1.33, 1.66 and 2.00. The noise level was calculated inside a region of the brain. This dataset was augmented to cover CNR levels from 0.33 to 3.00 – this augmented dataset is referred to as Hybrid Dataset One (HDS1).

For lower CNRs, the analysis of this dataset was confounded by the presence of a pre-existing structure adjacent to the square region of artificial activation (visible in Figure 47b). For this reason, a second hybrid dataset was created that displaced the region of artificial activation away from pre-existing structures. This dataset is referred to as HDS2 and is identical to HDS1 in all respects other than the location of the region of artificial activation (Figure 47). Even though the local noise level was somewhat lower in HDS2, for consistency with the original published dataset, the CNR levels were computed with same average noise level used in the original study.

### 5.3.1.2 Results for HDS1/HDS2

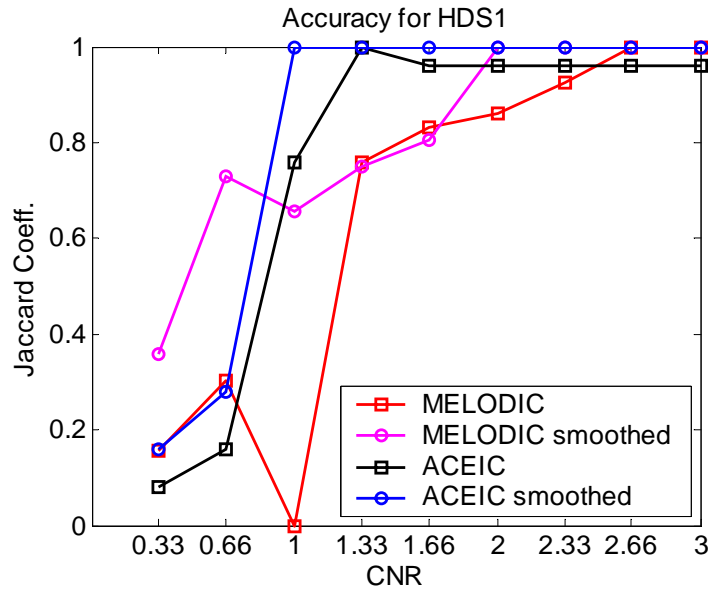
Since, for this segmentation task, the number of true-negatives (pixels outside the true region of activation that were not part of the solution segment) can overwhelm accuracy calculations, the Jaccard Coefficient ( $JC$ ) [40] was chosen as the measure of accuracy. The Jaccard Coefficient does not use the true-negative count in the accuracy computations. If the

number of true-positives, false-positives and false-negatives are denoted by  $TP$ ,  $FP$  and  $FN$  respectively, the Jaccard Coefficient is defined as

$$JC = \frac{TP}{TP + FP + FN} \quad (5.11)$$

For segmentation of HDS1 and HDS2 datasets, the ACEIC method was used with default options – segment size limit of 100 and seed selection based on spectral peaks. The accuracies of these segmentation results were compared with those of MELODIC [62] (Version ln(11), with default options), a publicly available implementation of Probabilistic ICA (PICA). Both methods were also repeated after temporal smoothing of the timecourses with Gaussian kernels (Full Width at Half Maximum of twice the scan repeat time, or TR).

Since MELODIC is designed to identify independent components and not segments, its solution to the segmentation problem was based on the independent component that exhibited maximum intersection with the true region of activation (see Figure 49). To highlight the segmentation capability of MELODIC, false-positives away from the true region of activation were ignored for MELODIC – with the inclusion of these false-positives, the MELODIC accuracies will be lower than reported. Figure 48 compares the accuracies for ACEIC and MELODIC for HDS1.



**Figure 48.** Comparison of accuracies of ACEIC and MELODIC for dataset HDS1.

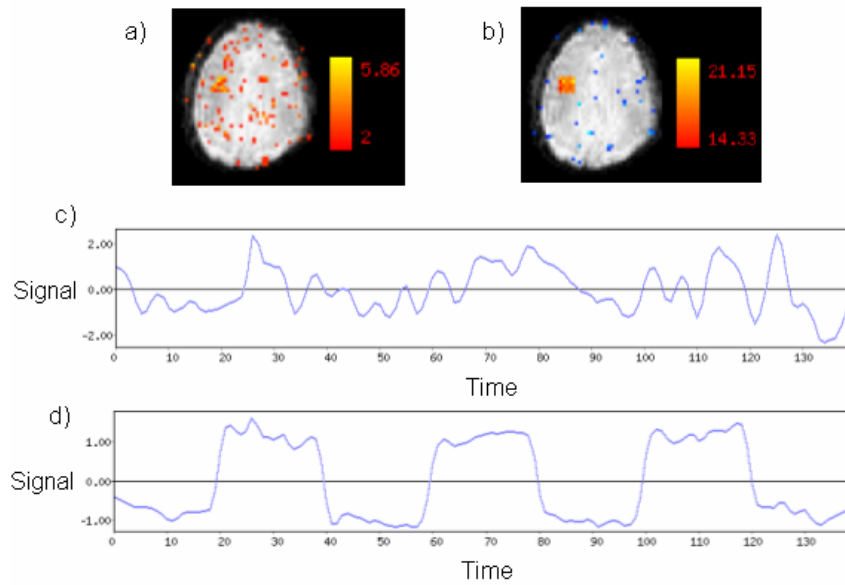
For HDS1 (Figure 48), the presence of the previously mentioned pre-existing structure near the true region of activation caused persistent over-segmentation by MELODIC – the



segmentation solutions included the pixels from the pre-existing structure. In one case (CNR=1 unsmoothed), MELODIC did not report a solution (no convergence during ICA estimation).

With temporally smoothed images, the ACEIC solution (blue line with round symbols in Figure 48) was accurate for CNRs 1 and above. However, without smoothing (black line with rectangular symbols in Figure 48), the ACEIC solution consistently included an extra pixel in the solution. This inaccuracy is caused by the fact that, without smoothing, the top of the plateau shown in Figure 46 may not be flat, leading to the selection of an incorrect threshold. For this reason, moderate amount of temporal smoothing is recommended for ACEIC.

In Figure 48, for CNR=0.66, the ACEIC method did not recover from the under-segmentation problem even with smoothing. While MELODIC (with temporal smoothing) outperforms ACEIC in this case, it is difficult to interpret the MELODIC solution due to the presence of clumps of false-positives away from the true region of activation (see Figure 49a). With the inclusion of all false-positives from MELODIC,  $JC=0.14$  for this case (CNR=0.66). It is also difficult to interpret whether the characteristic timecourse for this component (Figure 49c) matches the true activation timecourse (Figure 49d).

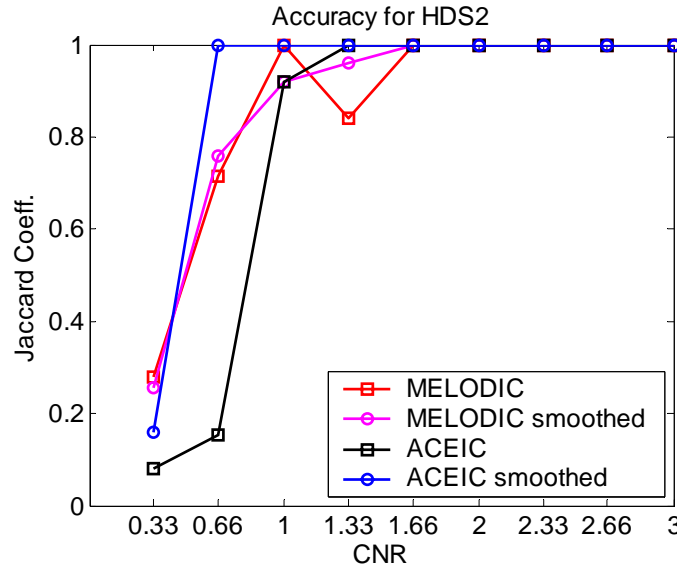


**Figure 49.** Examples of MELODIC solutions. a) Thresholded component map (showing z-scores) for CNR=0.66 (note false-positives). b) Thresholded Component map for CNR=3 (correct segmentation solution, no false positives). c) Timecourse associated with component shown for CNR=0.66. d) Timecourse associated with component shown for CNR=3.00.

For the HDS2 dataset (Figure 50), the accuracies of ACEIC and MELODIC were comparable for higher CNRs. For lower CNRs, ACEIC (with smoothing) exhibited better accuracy.

The effect of seed selection on the accuracy of ACEIC was explored by repeating the segmentation with all possible choices of seeds within the true regions of activation in HDS1 and HDS2. For CNR values where the default seeds yielded the correct solution, 5% of the alternate seeds produced incorrect solutions (overall standard deviations of  $JC$  were 0.14 and 0.06 for HDS1 and HDS2 respectively).

The ACEIC method was also evaluated for irregularly shaped regions of activation which were generated with truncated bivariate Gaussians (Figure 24) with randomly selected parameters (region size ranging from 12 to 28 pixels). The segmentation accuracies of ACEIC for these simulated regions of activation were similar to the results shown for HDS2 in Figure 50.

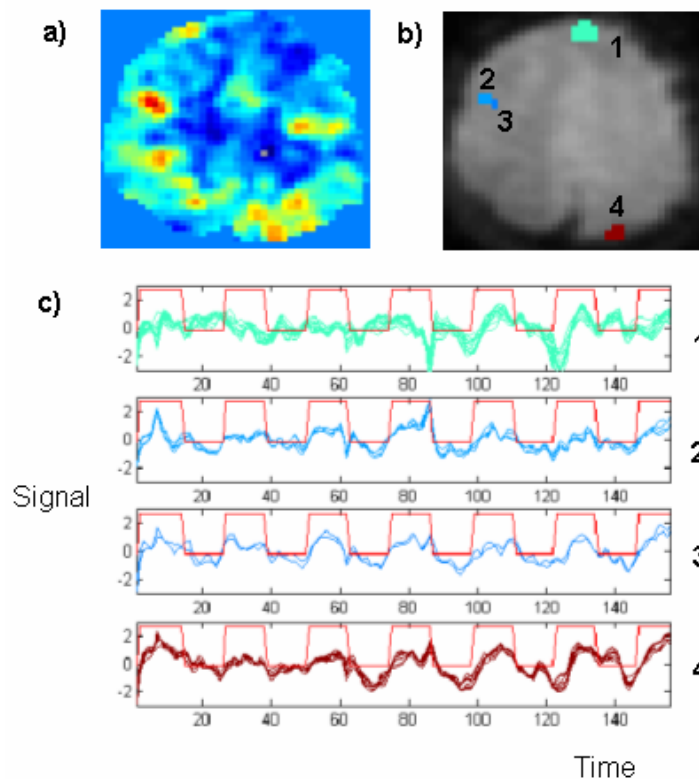


**Figure 50.** Comparison of accuracies of ACEIC and MELODIC for dataset HDS2.

### 5.3.2 Application to experimental data

To assess the usefulness of the ACEIC method for real-world data, fMRI data from a Visually Guided Saccade experiment was segmented with ACEIC. In this experiment, the subject performed seven 30-second blocks of a Visually Guided Saccade [75] task (see Chapter 3)

during the scan. For an axial slice of the brain, Figure 51a shows an activation image generated with voxel-by-voxel t-tests (section 2.2.2.2) comparing the signal levels during the saccade task to the signal levels during fixation. The durations of the saccade task are shown using a box-car representation in Figure 51c (red line). Figure 51b shows some of the functional segments identified by ACEIC – the corresponding time-courses are shown in Figure 51c. The region of strong activation in Figure 51a is also isolated by ACEIC (segments 2 and 3 in Figure 51b). While it would be desirable to isolate the functional unit into a single segment, this functional unit is reported as two segments due to noise in the signal (note the differences between the timecourses for segments 2 and 3 in the figure). This fracturing of functional units during segmentation requires that contiguous segments are merged before final segments are reported to the KDSf framework. Further experimentation with different distance metrics may provide a solution to this problem.



**Figure 51.** ACEIC segmentation of fMRI data from saccade experiment. a) T-test activation map showing regions of activation. b) A subset of segments identified by ACEIC. c) Timecourses associated with ACEIC segments.

Functional segmentation with ACEIC also isolated the presence of some artifacts in the data. Since the timecourses for segment 4 (Figure 51c) shares some of the strong peaks and

valleys with the timecourses for segment 1, and since they are located at opposite ends of the brain, it is likely that both segments are affected by head-motion of the subject. Thus, the relatively high level of activation associated with segment 4 (Figure 51a) incorporates contributions from head-motion. This shows that activation images cannot be accepted at face-value – motion-correction methods may be unable to remove all motion-related effects from the timecourses (these artifacts survived the motion correction procedure applied by FIASCO). As illustrated by this simple example, functional segmentation with ACEIC enables visualization of 4D datasets at the level of coherent groups of voxels – thus facilitating the detection of false-positives which may be missed by voxel-by-voxel inspection of timecourses.

## 5.4 DISCUSSION

Auto-threshold Contrast Enhancing Iterative Clustering (ACEIC) is a new method that demonstrates the feasibility of using region-contrast for functional segmentation of fMRI images. In this approach there are no parameters requiring tuning for individual images, as with conventional clustering methods. The main parameter required by the ACEIC method is the upper limit for region sizes, which can be chosen based upon available computational capacity and biological knowledge about extents of neural activation.

It was shown that maximization of region-contrast with greedy-agglomeration can lead to under-segmentation for regions of weak activation. The ACEIC method addresses this problem by a non-monotonic generalization of the greedy agglomeration approach that involves searching for the optimal region homogeneity threshold beyond which region-contrast is poor. In this work, the forward-difference operator was used to identify this optimal threshold – while this yields reasonable results, other approaches, such as searching for the widest unvarying portion of the plateau (see Figure 46), are also possible.

Evaluation of the ACEIC method with synthetic benchmark datasets indicates that the method can be accurate from CNR levels around 0.66 and higher. While the ACEIC method is more accurate than Probabilistic ICA for a range of activation levels, ACEIC is more computationally demanding – particularly for segmentation of full-brain images. During

exploratory analysis, it is possible to leverage the strengths of both methods by restricting the ACEIC seeds to the regions-of-interest isolated by PICA.

The exploration of the Visually Guided Saccade dataset indicates that functional segmentation with ACEIC can provide insights into the regions of activation identified by conventional methods of analysis. While co-activation of voxels implies coherence of timecourses for voxels within the region of activation, coherence of timecourses does not imply co-activation – other processes such as head-motion can also introduce coherence of timecourses. ACEIC provides a mechanism for isolation of voxels with coherent timecourses, which can then be analyzed to identify the causal process behind the observed coherence. Here it was shown that ACEIC can be useful for detection of coherence due to head-motion that survived the application of a standard motion-correction procedure. In future, it may be possible to utilize functional segmentation for automated detection and correction of such motion artifacts.

In summary, ACEIC provides a new coherence-based mechanism for functional segmentation that can be used with the KDSf framework. While ACEIC is demonstrated to be more accurate than PICA under controlled conditions, application of ACEIC to a real-world dataset uncovered the presence of spatial coherence from un-corrected head-motion along with spatial coherence from activation. The presence of these segments which are not related to activation complicates segment-based machine learning. It is difficult to eliminate these segments by comparing the timecourses for these segments with the task timecourse – as in this example, the t-score for these segments may be relatively high. Thus, while segmentation with ACEIC can help identify the presence of these artifacts, further improvements in motion-correction technology are needed before ACEIC can be applied for real-world functional segmentation.

## 6.0 MACHINE LEARNING WITH SCV

Although functional segmentation of real-world datasets is difficult with currently available tools, the spatial coherence principle can also be utilized for voxel-centric classification problems. In classification problems requiring discrimination between two groups of activation images (Statistical Parameter Maps), the spatial coherence principle refers to the homogeneity (coherence) of evidence for between-group differential activation within spatially contiguous voxels. For example, if all voxels in a large spatial cluster exhibit higher activation in the ‘disease’ group compared to the ‘normal’ group, such a cluster is likely to be a more reliable marker for ‘disease’ than a single voxel exhibiting similar between-group differential activation. Thus, regions of the brain that exhibit coherent differential activation are likely to be more useful for machine learning than regions with less coherence. In this Chapter, a new method for feature selection – Spatially Coherent Voxels (SCV) – is proposed for voxel-based classification problems, without requiring functional segmentation. It is shown that feature refinement with the SCV method can eliminate irrelevant features, leading to better classification accuracies than extant methods for feature selection.

Classification of activation images is complicated by the large number of voxels (features) in the individual activation images and the small number of subjects in the study (typically 15 – 20). Thus feature selection is an important consideration for automated classification tasks. Two dimensionality reduction techniques are commonly used with activation images – Principal Component Analysis (PCA) [2, 45] and  $k$ -best voxels (KBV) [44]. PCA is a standard method for dimensionality reduction that creates a smaller number of uncorrelated variables from linear combinations of correlated variables. In the KBV approach, the voxels are ranked by some criterion (e.g. activation level) and a certain number of the ‘best’ voxels are retained as features in the classification model. However, spatial relationships between the

voxels are not considered in the KBV approach – selected ‘best’ voxels may be contiguous or well-separated from each other. While the components identified by PCA incorporate spatial information, the maximum number of components is limited by the number of subjects in the dataset. Thus, small pockets of voxels that exhibit differential activation between groups of subjects may get incorporated into larger components which are not as effective for discrimination between the groups.

In this Chapter, the KBV approach is refined to exploit spatial relationships between the  $k$ -best voxels – the goal is to eliminate voxels that are less likely to be useful for classification. In this Spatially Coherent Voxels (SCV) approach, the  $k$ -best voxels are segmented into clumps of voxels based upon a spatial coherence criterion – two voxels are deemed to be coherent if they exhibit similar differential neural activation between the two groups of subjects (e.g. both voxels are under-activated in ‘disease’ subjects). Since larger clumps of coherent voxels are less likely to arise out of chance, filtering the  $k$ -best voxels based upon sizes of these clumps eliminates irrelevant features (voxels) from the classification model, leading to improved classification accuracies.

The next section discusses some of the prior work related to classification from activation images. In later sections, the SCV approach is described and applied to a classification problem from an fMRI study of Substance Use Disorder among adolescents. For this dataset, it is shown that the SCV approach yields better classification accuracies than PCA or KBV.

## 6.1 RELATED WORK

Several earlier studies have demonstrated the feasibility of automated classification of subjects based on functional neuro-images. Liow et al. [45] have used Positron Emission Tomography (PET) images to automatically classify HIV-positive patients from healthy controls. Ford et al. [2] have shown the feasibility of using fMRI activation maps to distinguish between patients and controls for Alzheimer’s disease, schizophrenia, and concussions. Zhang et al. [44] have studied the feasibility of distinguishing between healthy controls and patients with Substance Use Disorders (SUD). For Alzheimer’s disease, Pokrajac et al. [76] have explored classifications based upon spatial distribution of binary ROIs – without considering the activation strengths of

voxels. Also, Mitchell et al. [77] have used raw fMRI signal values (not activation images) to automatically classify cognitive states of subjects.

Dimensionality reduction is an important consideration for learning classification models from image datasets with small number of examples. Three main dimensionality reduction techniques have been considered in the past – PCA [2, 44, 45], KBV [44, 76, 77], and average over manually defined ROIs [77].

In PCA with activation images, a 3D image with  $m$  voxels is treated as an  $m$ -dimensional row vector and the data matrix  $X$  is constructed from  $n$  such rows, one for each subject. If the covariance matrix is of rank  $r$  (at most  $\min(n, m)$ ), the  $r$  principal components can be interpreted as ordered eigenimages that explain the variance in the image dataset (see Chapter 2). Projections of the original image vectors along these  $r$  basis vectors provide a reduced data matrix (of size  $n \times r$ ) for machine learning. Since the components are ordered by importance, the first few components in the model provide a coarse representation of the dataset. However, since  $r$  is limited by the number of subjects, none of the components may be sensitive to small pockets of differential activation between groups. Note that PCA is an unsupervised method of feature reduction – the class labels are not used for construction or selection of the components.

In the KBV approach, a number of voxels are selected based upon some criterion – such as  $k$ -most active [44, 76] or  $k$ -most active within a pre-defined ROI [44]. While the above approaches are unsupervised in the sense that class information from the training set is not used for feature selection, supervised methods can be more effective than unsupervised methods. This work employs a supervised approach, where the most discriminating voxels from the training set are retained in the classification model. The discriminating voxels can be identified from a group Statistical Parameter Map (SPM) [34] that employs a statistical test for each voxel to assess the level of between-group differential activation (in the training set). In a group SPM, each voxel location in the brain is assigned a statistical score (e.g. t-score) signifying the degree by which activation levels in one group differ from those in the other group. Applying the  $k$ -best voxels (KBV) feature selection strategy, it is possible to create classification models using the voxels with  $k$  highest t-scores in the group SPM of the training set.

However, the  $k$ -best approach does not consider spatial relationships between the selected voxels. Larger pockets of coherent differential activation are less likely to be caused by chance – this can be exploited to improve the feature selection process. In the SCV approach, rather than



using all the  $k$ -best voxels in the classification model, only those voxels that satisfy a spatial coherence criterion are utilized in the machine learning model. It is shown that this extra refinement can eliminate irrelevant features, leading to better classification accuracies compared to the conventional  $k$ -best approach.

## 6.2 METHODS

### 6.2.1 Dataset

An fMRI dataset from a study of Substance Use Disorders (SUD) among adolescents was used in this work to compare various feature selection approaches. In this dataset, the subjects were classified according to their score on a neurobehavioral disinhibition (ND) assessment. The ND score has previously been shown to be highly predictive of SUD and consists of measures of executive cognitive functions, affect modulation and behavioral control. Based upon these measures, 14 subjects in the study were assigned the ‘lowND’ label and 7 subjects were assigned the ‘highND’ label – these small sample sizes are typical of fMRI experiments. During the fMRI scan, all subjects performed an ‘anti-saccade’ task which required inhibition and reversal of reflexive saccadic movements of the eye. For analysis of data from a single subject, the signal level during the ‘anti-saccade’ task is compared with that from the ‘pro-saccade’ task (see Chapter 3) to isolate the regions of the brain activated during willful suppression of a reflexive response.

The scan data from the experiments were analyzed with FIASCO [78] and AFNI [79] to create neural activation maps for each subject – the t-score at each voxel in the 3D activation map provides a measure of the task-induced neural activation during inhibition (of the saccadic response). These activation maps were spatially normalized to a common coordinate system [29] to facilitate comparisons between brains of different shapes and sizes. The goal of the classification task is to automatically label subjects (as ‘highND’ or ‘lowND’) based upon the characteristics of these activation images.

### 6.2.2 Normalization

Summary statistics for t-scores in activation images showed substantial variation between subjects – for example, the maximum t-score within an activation image ranged from 4.6 to 15.5 between subjects. Preliminary analysis indicated that classification with raw t-scores as feature values was less accurate than classification with normalized t-scores. For normalization, the t-scores within each activation image were replaced with the corresponding percentile values within the particular activation image – thus, for each subject, the normalized values were close to one in regions of high neural activation. As described earlier, a group SPM was constructed from the normalized training set images – for each voxel, the t-score in the group SPM reflected the discriminative power of the individual voxel.

### 6.2.3 Feature Selection

With these normalized activation images, three methods of feature selection were explored – PCA,  $k$ -best voxels (KBV), and Spatially Coherent Voxels (SCV). For the PCA approach, the ‘snap-shot’ method [80, 81] was used for computational tractability. Instead of calculating an  $m \times m$  co-variance matrix, the ‘snap-shot’ method employs an alternate  $n \times n$  covariance matrix for computation of eigenvalues and principal components. Once the principal components were identified by PCA, the components were introduced as features in the machine learning models in the order of importance of the components (eigenvalues).

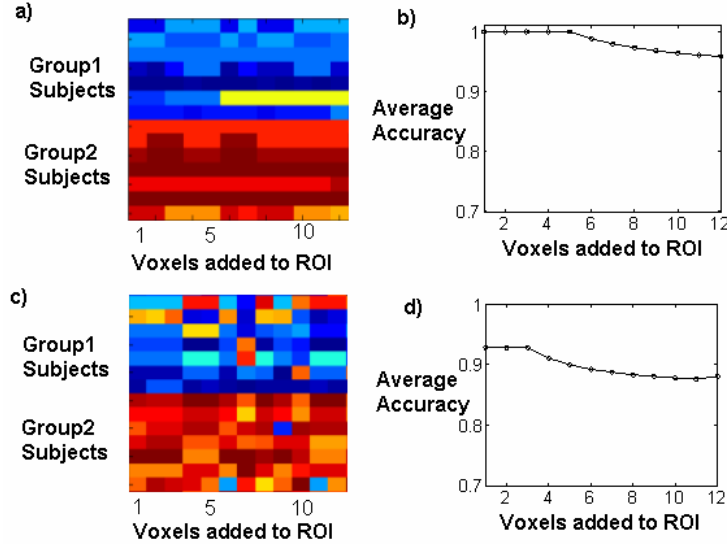
In the KBV approach,  $k$  most discriminating voxels from the training set were retained in the model. To identify the discriminating voxels, for each voxel location, activation values for the subjects in the training set were compared with a two-sample t-test. The voxels with the  $k$  highest t-scores were used as features in the classification models.

The SCV method built upon the KBV approach by retaining only a subset of the  $k$ -best voxels which satisfied a spatial coherence test. The  $k$ -best voxels were segmented based upon spatial coherence and only segments larger than a pre-specified size threshold were retained – these coherent segments are also referred to as regions of interest (ROI). While cluster-size thresholds have been used previously to increase statistical power for detection of activated voxels in an image from a single subject [82], use of coherent clusters across subjects for feature

selection has not been explored previously. The spatial coherence test is designed to eliminate small pockets of discriminating voxels, which are more likely to occur by chance. Once the discriminating voxels are filtered in this way, the filtered voxels are used as features in the same way as in the KBV approach. Details of the filtering employed by the SCV method are described next.

#### **6.2.4 Spatially Coherent Voxels (SCV)**

The discriminative coherence for an ROI is assessed by the Average Classification Accuracy (ACA) measure, which evaluates the degree to which voxels in the ROI consistently support between-group differences in neural activation. For each voxel in the ROI, the individual discriminative power of the voxel is assessed by the classification accuracy achieved by a classifier when only the between-subject activation levels at the particular voxel are considered in the classification model. The classification accuracy for each voxel is assessed by leave-one-out validation with a univariate Bayes classifier (with Gaussian likelihood). The ACA measure for an ROI is the mean of the individual classification accuracies for all the voxels in the ROI. The use of a univariate classifier in the ACA computation maintains the desired degree of coherence within the ROI – multivariate classifiers can yield high classification accuracies even in the presence of discordant voxels in the ROI. It should be noted that, while it is possible to construct ROIs by thresholding t-scores in the group SPM, that approach does not guarantee any particular classification accuracy from such ROIs – thus, it is not clear how the threshold should be chosen. Figure 52 illustrates activation levels from two different ROIs – one of the ROIs exhibit higher discriminative coherence than the other.



**Figure 52.** Example of spatially coherent voxels. a) Heat-map representation of activation values for a coherent ROI. b) For the coherent ROI, average accuracy (ACA) stays above threshold as the ROI is grown one voxel at a time. c) Activation values for less coherent ROI. d) Average accuracy quickly falls below threshold (0.9).

Segmentation with a greedy agglomeration strategy is used to identify ROIs with strong discriminative coherence (note that the  $k$ -best voxels are segmented, not individual activation images). Starting with a seed voxel, the neighboring voxel with the highest t-score in the group SPM is added to the ROI until the ACA measure computed for the current ROI definition falls below the desired threshold.

At the start of the ROI construction process, all the  $k$ -best voxels are placed in the set of currently available voxels  $A$ . A region (or segment)  $R$  is a non-empty subset of spatially connected voxels from  $A$  using the 26-connected neighborhood model for 3D images. If  $v$  and  $w$  are two voxels, the neighbor relationship is defined to be  $N(v, w) = 1$  iff  $v$  and  $w$  are 26-connected.

The exterior boundary  $E$  is the set of voxels outside the set  $R$  that have at least one neighbor voxel in the set  $R$ .

$$E = \{v \mid v \notin R, v \in A, w \in R, N(v, w)\} \quad (6.1)$$

If  $t_v$  is the t-score for voxel  $v$  in the group SPM of the training set, the ROI construction process starts with the voxel with the highest t-score in the set  $A$ .

$$R = \{p \mid p = \arg \max_{v \in A} \{t_v\}\} \quad (6.2)$$

Next, the exterior boundary  $E$  is computed for the current region definition ( $R$ ). The voxel in  $E$  with the highest t-score (voxel  $q$ ) is added to the region  $R$  (greedy agglomeration).

$$R = R \cup q, \quad q = \arg \max_{v \in E} \{t_v\} \quad (6.3)$$

If  $a_v$  is the leave-one-out classification accuracy for the training set subjects when only voxel  $v$  is retained in the Gaussian Bayes classification model, the Average Classification Accuracy  $a_R$  for the region  $R$  is the mean of  $a_v$  values computed from each of the voxels in the region.

$$a_R = \text{mean}_{v \in R} \{a_v\} \quad (6.4)$$

The external boundary  $E$  is recomputed and the region-growing step (Equation 6.3) is repeated as long as the  $a_R$  measure does not fall below a pre-specified coherence threshold (by default 0.9). When this coherence threshold can no longer be maintained (see Figure 52d), the region-definition is finalized and the voxels in  $R$  are removed from the available set  $A$ .

$$A = A - R \quad (6.5)$$

The segmentation process continues with another seed from the pool of currently available voxels (Equation 6.2). Once all the  $k$ -best voxels are segmented in this way, all regions (segments) below a size threshold  $\theta$  are discarded – the classification model is built with voxels inside regions that exceed this size threshold.

The SCV approach requires specification of three parameters –  $k$  (number of voxels to segment), the coherence threshold (default 0.9), and the ROI size threshold  $\theta$ . The optimal choice of  $\theta$  is dataset-specific and needs to be chosen empirically.

### 6.2.5 Classification

Feature refinement by spatial coherence (SCV) attempts to eliminate features which are irrelevant for classification – this is likely to improve classification accuracies. To test this hypothesis, a leave-one-out validation approach is employed. For a particular set of features (e.g. a particular value of  $k$  in KBV, or a number of principal components) and refinement parameters (e.g.  $\theta$ ), feature selection and model construction is performed with a particular subject left-out of the full dataset – the classification model is then tested on the subject left-out of the model.

This process is repeated  $n$  times with different subjects left-out of the training phase, where  $n$  is the number of subjects in the dataset. The average classification accuracy from the  $n$  repetitions is reported as the leave-one-out classification accuracy for the particular configuration of features and parameters. Leave-one-out cross-validation is a special case of multi-fold cross-validation that is appropriate for datasets with small number of examples.

Two different types of classification models (Chapter 2) are employed to test the hypothesis – a distribution-based approach and a distribution-free approach. The Gaussian Naïve Bayes (GNB) approach models the class-conditional likelihoods as independent Gaussian distributions and employs Bayes Rule for inference. The Support Vector Machines (SVM) approach identifies an optimal separating boundary to separate the two classes without modeling the distributions of the classes.

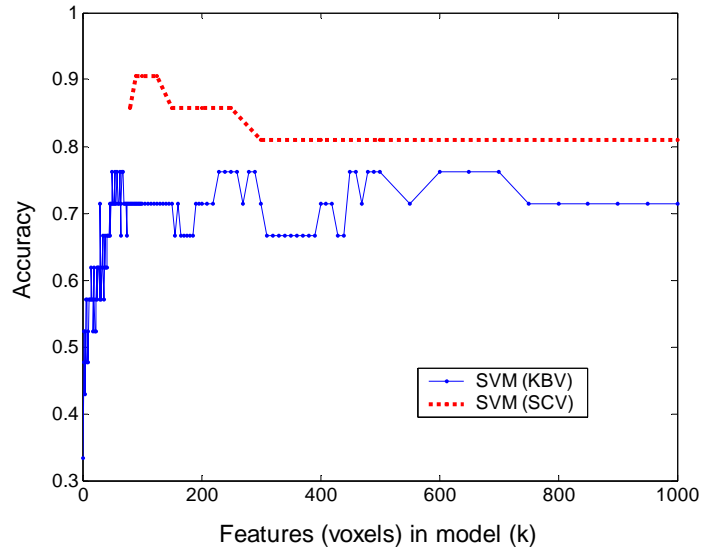
For these two types of classifiers (GNB and SVM), three feature selection methods (PCA, KBV and SCV) are compared. For each feature selection method, the parameter-space for the method is explored. For the PCA approach, features (components) are added to the classification model in the order of the corresponding eigenvalues. Thus, for  $n$  subjects in the training set,  $n$  classification models are tried – the first one with a single component and the last one with all the components. For the voxel based approaches (KBV and SCV),  $k$  is varied from 1 to 1000. For KBV, classification models are constructed with all these  $k$  voxels as features. For SCV, spatially coherent ROIs are identified from the  $k$ -best voxels and only voxels within ROIs larger than the specified size threshold ( $\theta$ ) are retained in the model. For each configuration of feature-selection parameters ( $n$ ,  $k$ , and  $\theta$ ), leave-one-out classification accuracies are computed.

### 6.3 RESULTS

Classification models were constructed from the ND dataset using three different feature selection methods – PCA, KBV and SCV. Two different types of classification models, Support Vector Machines (SVM) and Gaussian Naïve Bayes (GNB) were employed for comparing the accuracies from the three feature selection methods. The parameter-space for each of the feature selection methods was explored. For SCV, a grid-search over the possible values of  $\theta$  was used to determine the optimal value (yielding best leave-one-out accuracy). The best accuracies were

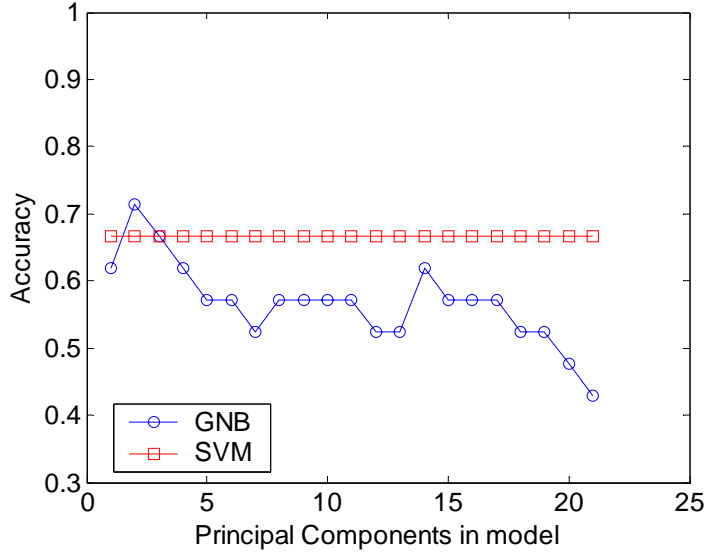
obtained for  $\theta$  values ranging from 15 to 25 – results for  $\theta=20$  are reported below, which corresponds to a ROI volume of  $20 \text{ mm}^3$  for this dataset.

The introduction of the spatial coherence requirement (SCV) in the feature selection process led to an improvement of classification accuracy. Figure 53 compares the accuracies of SVM classification models for the two voxel-based feature selection methods (SCV and KBV). As shown in the figure, elimination of the features (voxels) that did not satisfy the spatial coherence requirement improved classification accuracies – this was true for all values of  $k$ . When the number of retained voxels ( $k$ ) was below 80, the segmentation process did not consistently generate ROIs satisfying the spatial coherence requirements ( $a_R=0.9$  and  $\theta=20 \text{ mm}^3$ ). Thus, the SCV results are shown only for  $k$  higher than 80 voxels. For both SCV and KBV, the accuracies were lower when more than 1000 voxels were retained in the model.



**Figure 53.** ND dataset: Leave-one-out classification accuracies for SVM models with feature selection by KBV and feature refinement by SCV (with the same  $k$ ).

Leave-one-out classification accuracies for feature selection with PCA were lower than those from SCV and KBV (see Figure 54 and Table 12). While PCA has been reported to be effective in previous studies [2, 44], the method may not be suitable for small pockets of differentially activated voxels, as is the case for this dataset. The distribution of the voxels used in the leave-one-out iterations of the best SCV-based classification model (SVM and SCV with  $k=100$ ,  $a_R=0.9$  and  $\theta=20 \text{ mm}^3$ ) is shown in Figure 55. Over all the feature selection parameters ( $n$ ,  $k$ , and  $\theta$ ), the best leave-one-out classification accuracies are shown in Table 12.

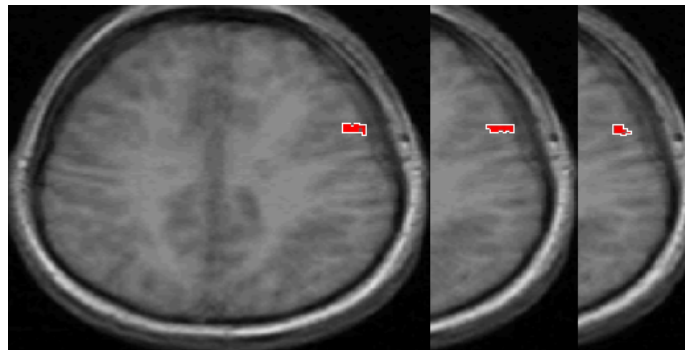


**Figure 54.** ND dataset: Leave-one-out classification accuracies for PCA-based feature selection, with different number of components in the model.

**Table 12.** ND dataset: Leave-one-out classification accuracies for the three feature selection methods – best accuracies achieved over respective parameter spaces (number of components,  $k$ , and  $\theta$ ).

	PCA	KBV	SCV
SVM	.667	.762	.904
GNB	.714	.667	.857

The spatially coherent voxels that were exploited by the best SCV-based classification configuration are shown in Figure 55 (overlaid on axial slices of the template brain [29]).



**Figure 55.** Location of spatially coherent voxels that were used in the best SCV-based classification configuration ( $k=100$  voxels,  $a_R=0.9$ , and  $\theta=20$  mm<sup>3</sup>).



## 6.4 DISCUSSION

This Chapter presented a novel feature selection method for classification from fMRI activation images. This method (SCV) uses a spatial coherence criterion to eliminate discriminating voxels without spatial support in fMRI image datasets. Features (voxels) without spatial support in the training set are more likely to be the product of chance and hence less likely to be generalizable to the testing set.

For a neurobehavioral disinhibition (ND) dataset, this feature refinement approach yielded better classification accuracies than the conventional approach of retaining the  $k$ -best voxels as features. Also, while PCA-based feature selection has been successfully employed for other datasets, results from this dataset indicate that PCA-based methods may not be sensitive to small regions of differential activation, particularly for datasets with small number of samples. While the benefit of the SCV approach has been demonstrated for this dataset, further evaluations with other datasets are needed to fully assess the potential of this method for real-world classification applications.

## **7.0 CONCLUSIONS AND FUTURE WORK**

Spatial coherence is the thread that binds the ideas presented in this work. Functional segmentation in the KDSf framework isolates brain voxels with coherent response to the task. Spatial coherence of timecourses is utilized by the ACEIC method for functional segmentation. The SCV method applies the spatial coherence principle to voxel-based feature selection. It has been demonstrated in this work that spatial coherence can be exploited to aid knowledge discovery and machine learning from fMRI datasets.

### **7.1 SPECIFIC FINDINGS**

The first hypothesis that feature construction via functional segmentation can improve classification accuracies from fMRI images has been validated for between-group differences in activation levels (Chapter 4). However, for between-group differences in sizes of regions of activation, PCA exhibited somewhat better classification accuracies than KDSf. Thus the value of functional segmentation for classification (claim 1) is established for some situations only. However, the superior interpretability of segmentation-based features may compensate for the small reduction in accuracies in these cases.

During the evaluation of the KDSf framework with simulated data, it was observed that the current approach of spatial smoothing of activation images to compensate for inadequate spatial normalization can lead to misinterpretation of the data. The activation levels in the smoothed images are dependent upon the sizes of the activated regions. Thus, with voxel-based knowledge discovery, differences in sizes of activated regions may be reported as differences in activation levels (see Figure 42). The KDSf framework does not suffer from this drawback.

The second hypothesis that spatial coherence of timecourses can be exploited for functional segmentation of fMRI images has been validated with benchmark datasets. The ACEIC method (Chapter 5) exploits the spatial coherence principle to segment fMRI images based upon similarity of timecourses. It has been demonstrated that the ACEIC method can achieve higher segmentation accuracies than Probabilistic ICA (claim 2).

The third hypothesis that spatial coherence information can also be helpful for classification from un-segmented images has been demonstrated by the SCV method for feature selection (Chapter 6). For a Substance Use Disorder dataset, the SCV approach achieved higher classification accuracies than traditional methods of feature construction (claim 3).

## 7.2 FUTURE WORK

While the potential benefits of functional segmentation has been established for knowledge discovery from fMRI data, real-world applications of this approach would require a practical method of functional segmentation. Even though the spatial coherence principle used by the ACEIC method showed promise with synthetic benchmark data, application of the technique to a real-world dataset showed that coherence may be caused by factors other than activation. That is, while activation implies spatial coherence, spatial coherence need not imply activation. Aside from head-motion, spatial coherence can be caused by physiological processes such as cardiac pulses. Thus, while coherence-based functional segmentation is feasible, application of this approach to real-world problems would require mechanisms for identifying the source of the spatial coherence. For example, variance of the activation response over time may indicate the presence of confounding factors other than neural activation.

As shown in chapter 5, functional segmentation with ACEIC indicates that even after application of motion-correction methods, some coherence-effects from head-motion of the subject may persist in the data. While this complicates the use of spatial coherence for functional segmentation, it may be possible to utilize the segmentation information to improve motion-correction techniques.

Another difficulty encountered with ACEIC is the fracturing of functional segments due to differences in timecourses reflecting the noisy nature of the data. While the correlation-based

distance-measure is commonly employed for comparing timecourses, it is somewhat sensitive to noise. A distance metric that is less sensitive to noise may alleviate this fracturing problem.

It may be possible to utilize the segmentation provided by the ACEIC method to achieve adaptive smoothing of fMRI images to improve the contrast-to-noise ratio (CNR). Usually, smoothing with isotropic Gaussian kernels is used for this purpose. However, the effectiveness of Gaussian smoothing kernels is dependent upon appropriate choice of ‘widths’ for the kernels. Averaging the timecourses within the ACEIC segments may lead to improvements in CNR, without requiring smoothing with Gaussian kernels.

While the ACEIC method is not yet ready for real-world use, the Spatially Coherent Voxels (SCV) approach to feature selection is suitable for immediate application to real-world datasets. The potential benefit of this approach has been demonstrated for the Substance Use Disorder dataset – evaluations with larger datasets will further establish its utility.

Finally, while the presentation of the KDSf framework in this work has been based upon characteristics of ROAs in the normalized coordinate system (template brain), it is also possible to explore the characteristics of ROAs in the original coordinate system of individual brains. This approach can avoid the loss of volumetric information associated with morphing of individual brains for spatial normalization. For this purpose, it is possible to segment the functional images in the original coordinate system and to spatially normalize the segmented images by applying the normalizing transformation to the segmented images. After registration of the ROAs in the normalized coordinate system, the resulting mapping between the ROAs and the functional units can be utilized to construct features based upon the characteristics of the ROAs in the original (un-morphed) coordinate system.

In conclusion, tools based on the spatial coherence principle add a new dimension to the toolkit for machine learning and automated knowledge discovery from fMRI data. Further development and validation of these tools will enhance the appeal of fMRI for clinical applications.

## APPENDIX

### GREEDY REGION REGISTRATION

Registration of regions of activations (ROAs) from different subjects is the problem of assigning labels to ROAs so that it is possible to examine the characteristics of *comparable* ROAs across subjects (see Chapter 4). A set of ROAs from different subjects in the same general anatomical neighborhood are assumed to correspond to a ‘functional unit’ of the brain – these ROAs are assigned a unique label for machine learning. While this registration can be done manually for small datasets, automated knowledge discovery from large datasets requires an automated solution. While many different approaches to the ROA registration problem are possible, a simple approach that does not require anatomical constraints is described below.

#### ROA Registration problem

The ROA Registration problem is similar to a clustering problem where the goal is to group ROAs based upon their spatial distribution in the brain images. To form clusters of ROAs, it is necessary to define distances between ROAs (an ROA is a set of connected voxels in an image).

#### Distance Between ROAs

Given two sets of voxels (regions)  $r1$  and  $r2$  that overlap (i.e. share some voxels), the distance  $d(r1, r2)$  between the two is defined in terms of the degree of overlap between the sets.

$$d(r1, r2) = 1 - \min\left\{\frac{|r1 \cap r2|}{|r1|}, \frac{|r1 \cap r2|}{|r2|}\right\} \quad (\text{A.1})$$

where  $r1 \cap r2 \neq \Phi$

If there is no overlap between  $r1$  and  $r2$ , the distance between the two regions is defined as the minimum Euclidean distance between the coordinates of voxels in the two sets.

$$d(r1, r2) = \min_{v1, v2} (\|v1, v2\|), \text{ where } r1 \cap r2 = \Phi, v1 \in r1, v2 \in r2 \quad (\text{A.2})$$

and, the distance between voxels  $v1$  and  $v2$  is the Euclidean distance

$$\|v1, v2\| = \sqrt{(x1 - x2)^2 + (y1 - y2)^2 + (z1 - z2)^2}$$

where  $v1 = [x1, y1, z1], v2 = [x2, y2, z2]$

### Problem Statement

Given a set of images and a set of ROAs in the images, find a set of ‘functional units’ (or just *units*) such that each unit claims at most one ROA from each image and each ROA belongs to one unit. Thus the problem is to find a mapping from each ROA in each image to a unit. The unit is similar to a cluster, with the above-mentioned constraint. The within-unit ‘error’ is defined as the sum of  $d(r1, r2)$  where  $r1$  and  $r2$  are ROAs from *different images* claimed by the same unit. The ‘width’ of a unit is the maximum distance  $d(r1, r2)$  between two ROAs  $r1$  and  $r2$  (from different images) in the same unit. The distance between two units is defined as the minimum distance  $d(r1, r2)$  where  $r1$  and  $r2$  are ROAs claimed by the two units.

Given the maximum allowable ‘width’ of any unit (a user-specified parameter  $W$ ), the objective of the ROA Registration problem is to find unit-definitions that minimize the total within-unit error, which is the sum of the within-unit errors for all units. To avoid a set of singleton units, it is also required that the distance between singleton units is greater than  $W$ . Thus, the ROA registration problem is similar to a clustering problem that tries to minimize sum-of-squared-error [83]. This is a combinatorial optimization problem and similar problems have been shown to be NP-hard.

### Greedy ROA Registration Algorithm

The Greedy ROA Registration (GRR) algorithm is similar to hierarchical clustering with complete linkage – in GRR the emphasis is on creation of functional units that are as inclusive of ROAs as possible (while satisfying the constraint on ‘widths’ of functional units). This approach attempts to ‘cover’ the data with larger functional units (clusters with more ROAs) than the traditional hierarchical clustering approach. More ROAs in a functional unit represents more subjects for whom the characteristics of the particular unit can be compared for machine learning.

In this algorithm, all the ROA-pairs (from different images) are first sorted by distance between the ROAs. The pair with the smallest distance is used to start a new *functional unit* (a

set of ROAs). The unit is greedily populated with ROAs as long as all member ROAs satisfy the maximum allowable distance between ROA-pairs in a unit ( $W$  the user-specified maximum ‘width’ of the unit). Once the unit runs out of eligible ROAs, a new unit is started. This greedy solution to the ROA Registration problem provides an approximate solution to the general combinatorial optimization problem. In Figure 56,  $r_i^n$  represents the  $i^{\text{th}}$  ROA in the  $n^{\text{th}}$  image.

### Algorithm GRR

Input1: A set of ROAs  $r_i^n$

Input2: 'Width' threshold  $W$  for functional units

Output: A set of functional units (a functional unit is a set of ROAs)

1. For all ROA pairs  $[r_i^m, r_j^n]$  where  $m \neq n$ , compute the distance between the ROAs  $d(r_i^m, r_j^n)$  (equations A.1,A.2).
2. Set  $d(r_i^m, r_j^n) = \infty$  when  $m = n$  so that ROAs from same image are not included in the same unit.
3. Add all ROAs  $r_i^n$  to the available set  $A$ .
4. While the set  $A$  is not empty
5. Find the available pair  $[r_k^m, r_l^n]$  with minimum pair-wise distance
$$[r_k^m, r_l^n] = \arg \min_{r_r^p, r_s^q} d(r_r^p, r_s^q) \mid r_r^p \in A, r_s^q \in A$$
6. If  $d(r_k^m, r_l^n) > W$
7. Place all ROAs in  $A$  in singleton functional units.
8. Empty the available set,  $A = \Phi$ . Stop.
9. end-if
10. Start a new functional unit  $U$  with this pair  $U = \{r_k^m, r_l^n\}$
11. Mark  $U$  as *unfinished*
12. Remove ROAs  $r_k^m$  and  $r_l^n$  from the available set  $A$ ,  $A = A - \{r_k^m, r_l^n\}$
13. While  $U$  is marked as *unfinished*
14. Find the *available* ROA  $r_o^q$  most eligible for merger with  $U$ 

$$r_d^t = \arg \min_{r_o^q \in A} d(r_o^q, r_l^n) \mid r_l^n \in U$$
15. If inclusion of  $r_d^t$  does not violate the width constraint
$$\max \{d(r_d^t, r_l^n) \mid r_l^n \in U\} \leq W$$
16. Add the ROA  $r_d^t$  to the unit  $U = U \cup \{r_d^t\}$
17. Remove  $r_d^t$  from the available set  $A = A - \{r_d^t\}$
18. else
19. save the ROAs in  $U$  as the next functional unit  $F_s, F_s = U$
20. Mark  $U$  as *finished*
21. end-if
22. end-while
23. end-while

**Figure 56.** Greedy ROA registration algorithm.



## BIBLIOGRAPHY

- [1] P. Golland, B. Fischl, M. Spiridon, N. Kanwisher, R. L. Buckner, M. E. Shenton, R. Kikinis, A. Dale, and W. E. L. Grimson, "Discriminative Analysis for Image-Based Studies," *Lecture Notes in Computer Science*, vol. 2488, pp. 508-515, 2002.
- [2] J. Ford, H. Farid, F. Makedon, L. A. Flashman, T. W. McAllister, V. Megalooikonomou, and A. J. Saykin, "Patient Classification of fMRI Activation Maps," in *Lecture Notes In Computer Science*: Springer-Verlag, 2003, pp. 58-65.
- [3] S. Ogawa, T. M. Lee, A. R. Kay, and D. W. Tank, "Brain magnetic resonance imaging with contrast dependent on blood oxygenation," *Proc. Natl. Acad. Sci. USA*, vol. 87, pp. 9868-9872, 1990.
- [4] M. Brett, I. S. Johnsrude, and A. M. Owen, "The problem of functional localization in the human brain.," *Nature Reviews, Neuroscience*, vol. 3, pp. 243-9, 2002.
- [5] A. Nieto-Castanon, S. S. Ghosh, J. A. Tourville, and F. H. Guenther, "Region of interest based analysis of functional imaging data," *Neuroimage*, vol. 19, pp. 1303-16, 2003.
- [6] J. F. Mangin, D. Riviere, O. Coulon, C. Poupon, A. Cachia, Y. Cointepas, J. B. Poline, D. Le Bihan, J. Regis, and D. Papadopoulos-Orfanos, "Coordinate-based versus structural approaches to brain image analysis," *Artif Intell Med.*, vol. 30, pp. 177-197, 2004.
- [7] G. M. Boynton, S. A. Engel, G. H. Glover, and D. J. Heeger, "Linear systems analysis of functional magnetic resonance imaging in human V1," *Journal of Neuroscience*, vol. 16, pp. 4207-4221, 1996.
- [8] M. E. Raichle, "Functional Neuroimaging: A Historical and Physiological Perspective," in *Handbook of Functional Neuroimaging of Cognition*, R. Cabeza and A. Kingstone, Eds.: MIT Press, 2001.
- [9] P. S. Mitra, V. Gopalakrishnan, and R. L. McNamee, "Segmentation of fMRI Data by Maximization of Region Contrast," presented at IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA), New York, 2006.
- [10] M. Jarmasz and R. L. Somorjai, "Exploring regions of interest with cluster analysis(EROICA) using a spectral peak statistic for selecting and testing the significance of fMRI activation time-series.," *Artif Intell Med.*, vol. 25, pp. 45-67, 2002.
- [11] F. Bloch, W. W. Hansen, and M. Packard, "Nuclear Induction," *Physics Review*, vol. 69, pp. 127, 1946.
- [12] E. M. Purcell, H. C. Torrey, and R. V. Pound, "Resonance Absorption by Nuclear Magnetic Moments in a Solid," *Physics Review*, vol. 69, pp. 37-38, 1946.
- [13] P. C. Lauterbur, "Image Formation by Induced Local Interactions: Examples Employing Nuclear Magnetic Resonance," *Nature*, vol. 242, pp. 190-191, 1973.
- [14] A. Kumar, D. Welte, and R. R. Ernst, "NMR Fourier Zeugmatography," *J. Magn. Reson.*, vol. 18, pp. 69-83, 1975.

- [15] J. Beutel, H. L. Kundel, and R. L. Van Metter, "Handbook of Medical Imaging," SPIE - The International Society for Optical Engineering, 2000.
- [16] M. S. Cohen, "Echo-planar imaging (EPI) and functional MRI," in *Functional MRI*, P. A. Bandettini and C. Moonen, Eds., 1998.
- [17] S. Clare, "Functional MRI : Methods and Applications," University of Nottingham, 1997.
- [18] E. Amaro, Jr. and G. J. Barker, "Study design in fMRI: basic principles," *Brain and Cognition*, vol. 60, pp. 220-32, 2006.
- [19] R. L. Buckner and J. M. Logan, "Functional Neuroimaging Methods: PET and fMRI," in *Handbook of Functional Neuroimaging of Cognition*, R. Cabeza and A. Kingstone, Eds.: MIT Press, 2001.
- [20] C. Windischberger, H. Langenberger, T. Sycha, E. M. Tschernko, G. Fuchsjager-Mayerl, L. Schmetterer, and E. Moser, "On the origin of respiratory artifacts in BOLD-EPI of the human brain.," *Mag. Res. Imaging*, vol. 20, pp. 575-82, 2002.
- [21] K. J. Friston, S. Williams, R. Howard, R. S. J. Frackowiak, and R. Turner, "Movement-related effects in fMRI time-series," *Magnetic Resonance in Medicine*, vol. 35, pp. 346-55, 1996.
- [22] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimisation for the robust and accurate linear registration and motion correction of brain images," *NeuroImage*, vol. 17, pp. 825-841, 2002.
- [23] S. Grootenboer, C. Hutton, J. Ashburner, A. M. Howseman, O. Josephs, G. Rees, K. J. Friston, and R. Turner, "Characterization and correction of interpolation effects in the realignment of fMRI time series," *Neuroimage*, vol. 11, pp. 49-57, 2000.
- [24] X. Hu, T. H. Le, T. Parrish, and P. Erhard, "Retrospective Estimation and Correction of Physiological Fluctuation in Functional MRI," *Mag. Res. Imaging*, vol. 34, pp. 201-212, 1995.
- [25] G. H. Glover, T.-Q. Li, and D. Ress, "Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR," *Magnetic Resonance in Medicine*, vol. 44, pp. 162-167, 2000.
- [26] G. K. Aguirre and M. D'Esposito, "Experimental design for brain fMRI," in *Functional MRI*, C. Moonen and P. A. Bandettini, Eds.: Springer, 1999.
- [27] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-B. Poline, C. D. Frith, and R. S. J. Frackowiak, "Statistical parametric maps in functional imaging: a general linear approach," *Hum. Brain Map.*, vol. 2, pp. 189-210, 1995.
- [28] J.-B. Poline, F. Kherif, and W. D. Penny, "Contrast and Classical inferences," in *Human Brain Function*, J. Ashburner, K. Friston, and W. Penny, Eds., Second ed: Academic Press, 2004, pp. 761--778.
- [29] J. Talairach and P. Tournoux, *Co-planar Stereotaxic Atlas of the Human Brain*. New York: Thieme Medical, 1988.
- [30] J. Ashburner and K. J. Friston, "Spatial Normalization," in *Brain Warping*, A. W. Toga, Ed.: Academic Press, 1999, pp. 27-44.
- [31] R. S. J. Frackowiak, K. J. Friston, C. D. Frith, R. J. Dolan, and J. C. Mazziotta, "Human Brain Function," Academic Press USA, 1997.
- [32] N. A. Lazar, B. Luna, J. A. Sweeney, and W. F. Eddy, "Combining brains: a survey of methods for statistical pooling of information," *Neuroimage*, vol. 16, pp. 538-550, 2002.
- [33] R. L. McNamee and N. A. Lazar, "Assessing the sensitivity of fMRI group maps," *Neuroimage*, vol. 22, pp. 920-31, 2004.

- [34] W. D. Penny and A. P. Holmes, "Random effect analysis," in *Human brain function*, R. S. J. Frackowiak, K. J. Friston, C. D. Frith, R. J. Dolan, C. J. Price, S. Zeki, J. Ashburner, and W. D. Penny, Eds. New York: Academic, 2003, pp. 843-850.
- [35] T. M. Mitchell, *Machine Learning*: McGraw Hill, 1997.
- [36] D. O. Hebb, *The Organization of Behavior*: John Wiley & Sons, 1952.
- [37] F. Rosenblatt, "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain," *Psychological Review*, vol. 65, pp. 386-408, 1958.
- [38] M. L. Minsky and S. A. Papert, *Perceptrons*. Cambridge, MA: MIT Press, 1969.
- [39] B. Widrow and M. A. Lehr, "30 years of Adaptive Neural Networks: Peceptron, Madaline, and Backpropagation," *Proc. IEEE*, vol. 78, pp. 1415-1442, 1990.
- [40] J. Han and M. Kamber, *Data Mining*: Morgan Kaufmann Publishers, 2001.
- [41] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, 1995.
- [42] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, Second Edition ed: Wiley, 2001.
- [43] C.-C. Chang and C.-J. Lin, "LIBSVM : a library for support vector machines," 2001 (<http://www.csie.ntu.edu.tw/~cjlin/libsvm>).
- [44] L. Zhang, D. Samaras, D. Tomasi, N. Volkow, and R. Goldstein, "Machine Learning for Clinical Diagnosis from Functional Magnetic Resonance Imaging," presented at IEEE Proceedings of CVPR, 2005.
- [45] J. S. Liow, K. Rehm, S. C. Strother, J. R. Anderson, N. Morch, L. K. Hansen, K. A. Schaper, and D. A. Rottenberg, "Comparison of voxel- and volume-of-interest-based analyses in FDG PET scans of HIV positive and healthy individuals," *J Nucl Med.*, vol. 41, pp. 612-21, 2000.
- [46] T. Yokoo, "Multivariate Statistical Analysis of Functional Neuroimaging Data," in *Mount Sinai School of Medicine*. New York, NY, 2004.
- [47] B. Thirion, G. Flandin, P. Pinel, A. Roche, P. Ciuciu, and J. B. Poline, "Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets," *Hum. Brain Map.*, vol. [Epub ahead of print] Nov 9, 2005.
- [48] P. Filzmoser, R. Baumgartner, and E. Moser, "A hierarchical clustering method for analyzing functional MR images," *Mag. Res. Imaging*, vol. 17, pp. 817-826, 1999.
- [49] C. Goutte, P. Toft, E. Rostmp, F. Nielsen, and L. K. Hansen, "On clustering fMRI time series," *NeuroImage*, vol. 9, pp. 298-310, 1999.
- [50] M. J. Fadili, S. Ruan, D. Bloyet, and B. Mazoyer, "A multistep Unsupervised Fuzzy Clustering Analysis of fMRI time series," *Human Brain Mapping*, vol. 10, pp. 160-178, 2000.
- [51] E. Dimitriadou, M. Barth, C. Windischberger, K. Hornik, and E. Moser, "A quantitative comparison of functional MRI cluster analysis," *Artificial Intelligence in Medicine*, vol. 31, pp. 57-71, 2004.
- [52] M. J. McKeown, S. Makeig, G. G. Brown, T. P. Jung, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski, "Analysis of fMRI data by blind separation into independent spatial components," *Hum. Brain Mapping*, vol. 6, pp. 160-188, 1998.
- [53] C. F. Beckmann and S. M. Smith, "Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging," *IEEE Transactions on Medical Imaging*, vol. 23, pp. 137-52, 2004.

- [54] V. Calhoun, T. Adali, L. Hansen, J. Larsen, and J. Pekar, "ICA of functional MRI data: an overview," presented at 4th International Symposium on Independent Component Analysis and Blind Signal Separation, Nara, Japan, 2003.
- [55] Y. Lu, T. Jiang, and Y. Zang, "Region growing method for the analysis of functional MRI data," *NeuroImage*, vol. 20, pp. 455-465, 2003.
- [56] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*: Wiley, 1990.
- [57] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, pp. 264-323, 1999.
- [58] S. C. Johnson, "Hierarchical Clustering Schemes," *Psychometrika*, vol. 2, pp. 241-254., 1967.
- [59] A. Hyvärinen and E. Oja, "Independent Component Analysis: Algorithms and Applications," *Neural Networks*, vol. 3, pp. 411-430, 2000.
- [60] A. J. Bell and T. J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129--1159, 1995.
- [61] V. D. Calhoun, T. Adali, G. D. Pearson, and J. J. Pekar, "Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms," *Human Brain Mapping*, vol. 13, pp. 43-53, 2001.
- [62] S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. J. Behrens, H. Johansen-Berg, P. R. Bannister, M. D. Luca, I. Drobnjak, D. E. Flitney, R. Niazy, J. Saunders, J. Vickers, Y. Zhang, N. D. Stefano, J. M. Brady, and P. M. Matthews., "Advances in functional and structural MR image analysis and implementation as FSL," *NeuroImage*, vol. 23, pp. 208-219, 2004.
- [63] D. H. Ballard and C. M. Brown, *Computer Vision*. Englewood Cliffs: Prentice-Hall, Inc., 1982.
- [64] S. A. Hojjatoleslami and J. Kittler, "Region growing: a new approach," *IEEE Transactions on Image Processing*, vol. 7, pp. 1079-84, 1998.
- [65] S. Tapert, A. D. Schweinsburg, V. Barlett, M. Meloy, S. Brown, G. Brown, and L. Frank, "Blood oxygen level dependent response and spatial working memory in alcohol use disordered adolescents," *Alcoholism: Clinical and Experimental Research*, vol. 28, pp. 1577-1586, 2004.
- [66] A. D. Schweinsburg, M. Paulus, V. Barlett, L. Killeen, L. Caldwell, C. Pulido, S. Brown, and S. Tapert, "An fMRI study of response inhibition in youths with a family history of alcoholism.," *Annals of the New York Academy of Science*, vol. 1021, pp. 391-394, 2004.
- [67] A. D. Schweinsburg, B. C. Schweinsburg, E. H. Cheung, G. G. Brown, S. A. Brown, and S. F. Tapert, "fMRI response to spatial working memory in adolescents with comorbid marijuana and alcohol use disorders," *Drug and Alcohol Dependence*, vol. 79, pp. 201-210, 2005.
- [68] R. E. Tarter, L. Kirisci, A. Mezzich, J. Cornelius, K. Pajer, M. Vanyukov, W. Gardner, and D. Clark, "Neurobehavior disinhibition in childhood predicts early age at onset of substance disorder," *American Journal of Psychiatry*, vol. 160, pp. 1078-1085, 2003.
- [69] C. Pierrot-Deseilligny, D. Milea, and R. M. Muri, "Eye movement control by the cerebral cortex," *Curr Opin Neurol.*, vol. 17, pp. 17(1):17-25, 2004.
- [70] B. Luna, K. R. Thulborn, D. Munoz, E. Merriam, K. Garver, N. Minshew, M. Keshavan, C. R. Genovese, W. Eddy, and J. Sweeney, "Maturation of widely distributed brain function subserves cognitive development," *Neuroimage*, vol. 13, pp. 786-793, 2001.

- [71] mathworks.com.
- [72] G. Tononi, A. R. McIntosh, D. P. Russell, and G. M. Edelman, "Functional clustering: identifying strongly interactive brain regions in neuroimaging data," *Neuroimage*, vol. 7, pp. 133-149, 1998.
- [73] J. G. Brankov, "Segmentation of Dynamic PET or fMRI Images Based on a Similarity Metric," *IEEE Transactions on Nuclear Science*, vol. 50, pp. 1410-1414, 2003.
- [74] K. J. Friston, O. Josephs, E. Zarahn, A. P. Holmes, S. Rouquette, and J.-B. Poline, "To smooth or not to smooth? Bias and efficiency in fMRI time-series analysis," *Neuroimage*, vol. 12, pp. 196-208, 2000.
- [75] B. Luna, K. R. Thulborn, M. H. Strojwas, B. J. McCurtain, R. A. Berman, C. R. Genovese, and J. A. Sweeney, "Dorsal cortical regions subserving visually guided saccades in humans: an fMRI study," *Cerebral Cortex*, vol. 8, pp. 40-7, 1998.
- [76] D. Pokrajac, V. Megalooikonomou, A. Lazarevic, D. Kontos, and Z. Obradovic, "Applying spatial distribution analysis techniques to classification of 3D medical images," *Artificial Intelligence in Medicine*, vol. 33, pp. 261-80, 2005.
- [77] T. M. Mitchell, R. Hutchinson, M. Just, R. Niculescu, F. Pereira, and X. Wang, "Classifying Instantaneous Cognitive States from fMRI data," *Proceedings of The American Medical Informatics Association*, 2003.
- [78] W. F. Eddy, M. Fitzgerald, C. R. Genovese, A. Mockus, and D. C. Noll, "Functional image analysis software - computational olio.," presented at Proceedings in Computational Statistics, Heidelberg, 1996.
- [79] R. W. Cox, "AFNI: software for analysis and visualization of functional magnetic resonance neuroimages," *Comp. Biomed. Res.*, vol. 19, pp. 162-173, 1996.
- [80] L. Sirovich and R. Everson, "Management and analysis of large scientific datasets," *Int'l J. of Supercomputer Applications*, vol. 6, pp. 50-68, 1992.
- [81] L. Sirovich, "Turbulence and the dynamics of coherent structures, pt. i: Coherent structures, pt. ii: Symmetries and transformations, pt. iii: Dynamics and scaling," *Quarterly of Applied Mathematics*, vol. XLV, pp. 561-590, 1987.
- [82] S. D. Forman, J. D. Cohen, M. Fitzgerald, W. F. Eddy, M. A. Mintun, and D. C. Noll, "Improved Assessment of Significant Activation in Functional Magnetic Resonance Imaging (fMRI): Use of a Cluster-Size Threshold," *Magnetic Resonance in Medicine*, vol. 33, pp. 636-647, 1995.
- [83] C. J. Veenman, M. J. T. Reinders, and E. Backer, "A maximum variance cluster algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1273-1280, 2002.